



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.



Stance Characterization and Detection On Social Media

Abeer AlDayel

Doctor of Philosophy

Institute for Language, Cognition and Computation

School of Informatics

University of Edinburgh

2021

Abstract

Stance detection refers to the task of identifying a viewpoint as either supporting or opposing a given topic. The current research on socio-political opinion mining on social media is still in its infancy. Most computational approaches in this field are limited to the independent use of textual elements of a user's posts from social factors such as homophily and network structure. This thesis provides a thorough study of stance detection on social media and assesses various online signals to identify the stance and understand its association with the analysed topic. We explore the task of detecting stance on Twitter, which is a well-known social media platform where people often express stance implicitly or explicitly.

First, we examine the relation between sentiment and stance and analyse the interplay between sentiment polarity and expressed stance. For this purpose, we extend the current SemEval stance dataset by annotating tweets related to four new topics with sentiment and stance labels. Then, we evaluate the effectiveness of sentiment analysis methods on stance prediction using two stance datasets.

Second, we examine the multi-modal representation of stance on social media by evaluating multiple stance detection models using textual content and online interactions. The finding of this chapter suggests that using social interactions along with other textual features can improve the stance detection model. Moreover, we show how an unconscious social interaction can reveal the stance.

Next, we design an online framework to preserve users' privacy concerning the implicitly inferred stance on social media. Thus, we evaluate the effectiveness of the two stance obfuscation methods and use different stance detection models to measure the overall performance of the proposed framework.

Finally, we study the dynamics of polarized stance to understand the factors that influence online stance. Particularly, we extend the analysis of online stance signals and examine the interplay between stance and automated accounts (bots). Furthermore, we pose the problem of gauging the bots' effect on polarized stance through a sole focus on the diffusion of bots on the online social network.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Abeer Aldayel)

بِسْمِ اللّٰهِ الرَّحْمٰنِ الرَّحِیْمِ

¹ وَقُلْ رَبِّ زِدْنِيْ عِلْمًا

This dissertation is dedicated to my family, friends, and all of my former students.

¹”My Lord, increase me in knowledge.” (TaHa. Quran verses 114) [سورة طه (١١٤)]

Acknowledgments

I have never had more people to thank. So many individuals have generously contributed their time, support and knowledge to this thesis. For this reason, I would humbly like to offer my gratitude to all those who have made this dissertation possible. I owe enormous debt of gratitude to my supervisor for the ongoing feedback, support and encouragement. The consistent pushing of my thinking challenged me in all the right ways. I thank Dr. Walid Magdy for the opportunity to learn and extend my scope of knowledge through this PhD journey. I am also thankful to my annual review committee members for their feedback and helping me question assumptions and view the research from multiple perspectives. Without the support and experience of my peers at the SMASH group, this thesis would not have been possible. I express tremendous gratitude towards the first cohort of the SMASH team, which started in September 2018, where we grew together and navigated the research field with a mix of eagerness and joy. I am thankful for the opportunity to be surrounded by extraordinary team members.

My deepest gratitude goes to those who have been a source of inspiration in completing this dissertation. Thanks to Dr.Kareem Darwish for his amazing effort and cooperation in conducting stance detection tutorial. My appreciation also goes to Dr. Tamer Elsayed and his team for the opportunity to work together and further navigate the stance study in an interesting research project.

I am very fortunate and grateful to King Saud University and Saudi Arabia Cultural Bureau for enabling and financing my studies in the University of Edinburgh and for giving me the chance to participate in many scientific venues.

Not least of all, I owe a lot to my family for their unlimited support and their unwavering belief that I can achieve so much. A special gratitude to my loving parents, Ibraheem Aldayel and Hessah Alomar, whose words of encouragement and push for tenacity still ring in my ears. I am grateful to my sisters, Ghada and Reema, and my brothers, Nawaf, Abdulrahman, Ahmad and Abdulmajeed, for their unwavering support, patience and love.

Alhamdulillah rabbi alamin

الحمد لله رب العالمين

Table of Contents

1	Introduction	1
1.1	Overview	1
1.2	Research questions	3
1.3	Contributions	4
1.4	Outline	5
1.5	Publications and research outcomes	7
2	Background and related work	9
2.1	Overview	9
2.2	Stance definition	11
2.3	Stance detection tasks	13
2.3.1	Stance detection according to target	13
2.3.2	Stance detection vs stance prediction	16
2.4	Stance modeling on social media	19
2.4.1	Content features	19
2.4.2	Network features	21
2.4.3	Comparative analysis of stance modeling	22
2.5	Stance detection algorithms	23
2.5.1	Supervised learning	23
2.5.2	Unconstrained supervised learning	25
2.5.3	Unsupervised learning	27
2.5.4	Best performing algorithm	27
2.6	Stance detection applications	29
2.6.1	Analytical studies	29
2.6.2	Applications related to social-based phenomena	30
2.6.3	Veracity checking applications	31
2.7	Stance Detection Resources	32

2.8	Current trends	34
2.9	Summary	35
3	Evaluation of sentiment as stance	37
3.1	Introduction	37
3.2	Tasks definitions	39
3.3	Association between sentiment and stance in literature	40
3.3.1	Sentiment as stance	40
3.3.2	Sentiment as a proxy for stance	41
3.4	Stance and sentiment datasets	42
3.4.1	SemEval stance dataset	42
3.4.2	Context-dependent dataset	43
3.5	Correlation between sentiment and stance	43
3.5.1	Agreement between sentiment and stance labels	43
3.5.2	Analysis of the textual patterns	45
3.5.3	The association between sentiment and stance	46
3.6	Sentiment as predictor of the stance	47
3.6.1	Sentiment models	47
3.6.2	Evaluation metric	48
3.6.3	Performance of sentiment analysis with stance	49
3.7	Summary	51
4	Possible factors for stance detection on social media	53
4.1	Introduction	53
4.2	Related work	56
4.3	Stance Detection Methodology	57
4.3.1	User vs Tweet features for Stance Detection	57
4.3.2	Feature Extraction	58
4.3.3	Stance Detection Model	59
4.4	Experimental Setup	60
4.4.1	Network Features to Detect Unexpressed View	60
4.4.2	Data Collection	60
4.4.3	Baselines and Evaluation	61
4.5	Results	63
4.5.1	Stance Detection Results	63
4.5.2	Performance Discussion	65

4.6	Feature Analysis	67
4.6.1	Similarity between Networks	67
4.6.2	Which Network Features Reveal the Stance?	69
4.6.3	The Context of the Features	71
4.7	Discussion	73
4.7.1	What factors predict the stance?	73
4.8	Summary	75
5	Stance obfuscation	77
5.1	Introduction	77
5.2	Related work	79
5.3	Obfuscation framework	80
5.3.1	Stance obfuscation feature space	81
5.3.2	Obfuscation methods	81
5.4	Experimental Setup	82
5.4.1	Dataset	82
5.4.2	Stance detection models	82
5.5	Results and discussion	83
5.5.1	Effectiveness of obfuscation methods	83
5.5.2	Additional paradigm	85
5.6	Summary	86
6	Characterizing the role of bots' in polarized stance on social media	87
6.1	Introduction	88
6.2	Related Work	90
6.2.1	Twitter policy on bots	90
6.2.2	Bots' role in social networks	91
6.3	Data collection	93
6.3.1	Stance-detection datasets	93
6.3.2	Collecting users' online networks	94
6.4	Assessing the role of social bots	94
6.4.1	Stance detection classifier	95
6.4.2	Extracting the most influential features on stance	96
6.4.3	Identification of bot accounts	97
6.5	Results and Analysis	98
6.5.1	The distribution of bot scores of the most influential accounts	98

6.5.2	The role of social bots on stances	100
6.5.3	Magnitude of the bots' role	101
6.5.4	Properties of the influential bot accounts	103
6.5.5	The context of the influential bots	105
6.6	Inspecting the deleted accounts	107
6.7	Verifying the bot/non bot accounts	110
6.8	Discussion	111
6.8.1	Bot and human effect on stances	111
6.8.2	The link between bots and supporting versus opposing stances	112
6.8.3	Bots' link to stance based on the interactions type	113
6.8.4	Implications	114
6.8.5	Limitations	115
6.9	Summary	115
7	Conclusion	117
7.1	Thesis contributions and findings	117
7.2	Limitations and future directions	120
	Bibliography	123
	Appendices	149
	Appendix A Background	149
A.1	List of recent stance prediction and detection work	149
A.2	Datasets for stance classification and prediction tasks	155
	Appendix B Stance obfuscation	159
B.1	Survey results	159
	Appendix C Socialbots and stance	161
C.1	Distribution of accounts scores on topic level	161
C.2	Distribution of bots on topic level	162
C.3	Chi-squared test for accounts distributions	163

List of Figures






1.1	Thesis organization and the related research questions.	6
2.1	The stance triangle, adapted from (Du Bois, 2007)	11
2.2	Stance representation in social media	19
3.1	The distribution of sentiment and stance with respect to each topic.	44
3.2	Distribution of sentiment per a given stance.	45
3.3	Jaccard similarity of the top N-most frequent words between sentiment and stance.	46
3.4	Tweets with matching and mixed stance and sentiment.	46
4.1	Confusion matrices for the best three vs two classes prediction models.	65
4.2	Similarity between CN, IN and DM in users dataset.	68
4.3	Similarity between CN, IN and DM for (In-favor and Against) stances with respect to the top features.	69
5.1	Effectiveness of stance obfuscation on four stance detection models—SVM, LR, NB and CNN, in comparison with the random stance detection model.	84
6.1	Botometer score distribution of the top 1000 accounts that are predictive to stance for both networks.	99
6.2	Distribution of social bots for each topic in the top 1,000 most predictive accounts for polarized stances using direct interaction (IN) and indirect exposure (EXP) features.	100

6.3	The percentage of each account type (X-axis) in the top N (Y-axis) influential accounts in predicting the Against/Favor stances in direct interactions (IN) and indirect interactions (EXP).	102
6.4	Distribution of social bots types for each topic in the top 1,000 most predictive accounts for polarised stances using direct interaction (IN) and indirect interaction (EXP).	104
B.1	Participants' ability to identify the stance indicated by different features.	160
B.2	The degree to which participants feel the need to avoid revealing their stance.	160
C.1	The distribution of accounts scores on the top 1,000 influential accounts from direct interactions (IN) in predicting the Against/Favor stance (Topic level).	161
C.2	The distribution of accounts scores on the top 1,000 influential accounts from in direct interactions (EXP) in predicting the Against/Favor stance (Topic level).	161
C.3	The distribution of bots on the top 1,000 influential accounts from the direct interactions (IN) in predicting the Against/ Favor stance (Topic level).	162
C.4	The distribution of bots on the top 1,000 influential accounts from indirect exposure (EXP) in predicting the Against/ Favor stance (Topic level).	162

List of Tables

2.1	Examples of stance from SemEval2016 stance dataset	12
2.2	A comparison of stance detection models using network , content and both as features.	24
2.3	Comparing the Stance detection models on SemEval stance dataset.	28
3.1	Number of tweets for each topic.	43
3.2	Sample of tweets illustrating the sentiment (Sent) polarity of the expressed stance (Stan). The examples are collected from SemEval stance and CD datasets.	48
3.3	Sentiment and stance prediction models on SemEval stance dataset. "GN" indicates general out-domain trained model and "CS" is for custom in-domain trained models.	50
3.4	Sentiment and stance prediction models on CD dataset. "GN" indicates general out-domain trained model and "CS" is for custom in-domain trained models.	50
4.1	List of feature sets examined in our experiments with their description.	59
4.2	Number of tweets used for training and testing with respect to Semeval 2016 topic. The number of unique users authored the tweets are shown in brackets.	60
4.3	Stance detection performance using different set of features using SVM classifier trained on three classes. F-Score (%) is reported on the SemEval stance detection task for each topic and overall. The set of features are categorized into three sets, namely, Interaction Network (IN), Preference Network (PN), and Connection Network (CN).	64

4.4	Stance detection performance using different set of features using <i>binary</i> SVM classifier. F-Score (%) is reported on the SemEval stance detection task for each topic and overall. The set of features are categorized into three sets, namely, Interaction Network (IN), Preference Network (PN), and Connection Network (CN).	64
4.5	The result of baseline linear SVM model when combining both text and network features. Model (A) and (B) shows the result when trained on three and two classes, respectively.	65
4.6	Top features extracted from the best model in each case and trained on two classes, CN_{FR} , $IN_{@}$, $PN_{@}$	69
4.7	Top features extracted from the best model in each case and trained on two classes, IN_{DM} , PN_{DM}	70
4.8	Sample of tweets and the context of IN and PN in relation with stance and topic.	72
5.1	Distribution of tweets in the dataset for the five topics.	82
5.2	F1 scores of stance detection algorithm before the hiding process.	83
6.1	The number of tweets per topic in the SemEval and Events datasets with the number of unique users who authored the tweets shown in brackets. The total number of accounts users interacted with ($IN_{@}$) and followed (CN_{FR}) for each topic.	95
6.2	The average F1-score for stance detection on the seven topics in our two datasets.	96
6.3	Distribution of bots and human based on followers. The Ultra-famous accounts > 10,000 followers; The famous accounts are those with number of followers ranging between 10,000 and 1,000; The normal accounts < 1,000 followers.	105
6.4	Sample of tweets and the context of social bot interactions in relation to stance and topic.	106
6.5	Top bot accounts in indirect interactions for each stance towards the seven topics.	107
6.6	The number of deleted accounts and the expected bots in the top 100 influential accounts on stance prediction.	109

6.7	Sample of tweets that from accounts that interacted with deleted accounts in the top 100 features of (IN). We used "X" to mask some users accounts and hide sensitive content.	110
6.8	Sample of verified accounts with explanation from Bot-Detective tool.	111
A.1	Work in stance prediction	150
A.2	Work in stance classification	155
A.3	Publicly available data-sets with stance annotations for stance classification in social media (in chronological order). Sources:  Twitter and  Reddit. Types "C": Claim-based, "T": Target based, and "MT": Multi related targets. The  indicates a multi-model dataset that contains contextual data along with the text. In stance annotation S, inputs "T": Target and "C": Claim. . . .	156
A.4	Publicly available data-sets with stance annotations for stance predictions (in chronological order). Sources:  : Twitter and  : News comments or online forum data.	157
C.1	Chi-squared test for accounts distributions between IN and EXP bot accounts. * $p_i < 0.05$, ** $p_i < 0.01$, *** $p_i < 0.001$	163

Chapter 1

Introduction

1.1 Overview

Nowadays, social media platforms constitute a major component of an individual's social interactions. People rely on these tools as the main source of news, as well as to connect to the world and get instant updates (Newman, 2011). They have a major beneficial side that allows individuals to explore various aspects of an emerging topic, express their own viewpoint, get instant feedback and examine the public's views. The huge dependency of people on these platforms as the main source of communication has allowed researchers to study the online public stance towards various topics.

Stance is defined as an expression of the speaker's standpoint and judgment towards a given proposition or object (Biber and Finegan, 1988). (Kockelman, 2004) described it as a semiotic means as stance is usually not explicit in the conversation and the description of how an individual feels can be used to attribute personal value to an object (Du Bois, 2007).

Stance has been used in various studies as a mean to link linguistic forms and social identities, which has the capability to better understand the background of people with a polarised stance (Bassiouney, 2015). Consequently, the literature on stance detection has focused on analysing the debates conducted on online forums (Lin et al., 2006; Somasundaran and Wiebe, 2009), which can be distinguished from the recent studies that have focused on social media platforms, especially Twitter, because the former has a clear single context in comparison to social media platforms. In online forums, users debate in the form of a thread discussion in which the information flow is usually focused on the topic (Belkaroui, Faiz, and Elkhelifi, 2014). In contrast, social media discussions on a given topic are more scattered; however, sometimes they can

be linked over a given hashtag (Mohammad et al., 2016b). Stance detection plays a major role in analytical studies conducted to evaluate public opinion on social media towards an event or topic. The stance detection process is also known as perspective (Beigman Klebanov, Beigman, and Diermeier, 2010; Elfardy, 2007) or viewpoint (Zhu, He, and Zhou, 2019; Trabelsi and Zaïane, 2018) detection, where perspective is identified by expressing stance towards an object of a controversial topic (Elfardy, 2007).

One of the earliest initiatives to promote stance detection on social media is the "SemEval 2016 stance detection" shared task, which introduced a stance-specific benchmark dataset to help in evaluating stance detection on Twitter (Mohammad et al., 2016b). In addition, a new wave of stance detection applications has been triggered to handle some issues that have infected the social media lately, such as fake news and rumours (Derczynski et al., 2017a,b; Aker et al., 2017), where the stance towards claims in a piece of news is used as a key feature for validating the credibility of the news.

Recent attempts have been made to gauge the online stance by identifying the pure polarity towards an event using sentiment analysis. Such modelling might be sub-optimal in representing the support stance. Therefore, various stance detection methods have been used to robustly identify viewpoints towards a topic (Magdy et al., 2016; Darwish, Magdy, and Zanouda, 2017b; Dey, Shrivastava, and Kaushik, 2018; Vijayaraghavan et al., 2016). Stance detection is the task of inferring whether a viewpoint towards a given topic or entity is supportive or against (Biber and Finegan, 1988). Recently, this task has attracted considerable attention due to its value as a social sensing method and as a downstream task for studying rumours and fake news on social media (Ma, Gao, and Wong, 2018). Traditional opinion methods, which rely on surveys and polls, have been proved to be limited in terms of cost and time. Consequently, the studies in this realm have used different methods for stance identification and socio-political opinion mining. The study conducted by (Murphy, Hill, and Dean, 2014) showed the correlation between opinion mining from social media and polls and surveys. In addition, previous studies have focused on textual cues to detect individual stances (Augenstein, Vlachos, and Bontcheva, 2016; Augenstein et al., 2016; Siddiqua, Chy, and Aono, 2018). There is a noticeable relation between individual behaviour and the detection of individual stances (Du Bois, 2007). This behavioural characteristic can be inferred from the vocabulary choice and online interactions within social media platforms.

1.2 Research questions

Stance detection on social media aims to determine whether people are in favor of or against a specific topic or event. The growing interest in employing user-generated information on social media to extract and determine individuals' stances has afforded the stance detection task substantial attention. The studies in this realm tend to use various online signals, ranging from textual content to online network connections of these platform users. In this thesis, we study stance modelling on social media, motivated by the arguments made by (Kockelman, 2004): 'subjectivity in language' is not an issue in stance studies; rather, stance may be interpreted as the intersection of a crosslinguistic account of the evaluated events and understandings of an individual's contribution. Moreover, (Du Bois, 2007) explained that the stance-taking process is affected by personal opinions and non-personal factors, such as cultural norms and social aspects. Therefore, we begin by assessing the sentiment polarity of an expressed stance and evaluate the effectiveness of sentiment methods for stance detection. Then, we examine the online interactions that can predict an individual's stance. To address this challenge, we design a stance detection model by using different groups of online social signals as features. Thus, we address the following four research questions in this study:

- RQ1: Can sentiment polarity be used to identify the stance towards an event? How are sentiment polarity and stance related?
- RQ2: To what extent do online interactions predict an individual's stance?
- RQ3: What are the minimal number of online features that can be injected or removed from an individual social media activity that can mislead stance models from predicting the stance?
- RQ4: How can the stance detection model be used to evaluate the interplay of bots with online stance?

To answer these questions, we use two stance datasets. First is the SemEval 2016 stance dataset, which contains five topics covering political, social, and religious domains. Second is a new dataset that we create by introducing the context-based dataset (CD dataset). In this dataset, we follow the annotation guideline same as that of the SemEval 2016 stance dataset, where we annotate each tweet with sentiment and stance. This new dataset contains four topics covering multiple domains. Further details about

these two datasets are mentioned in Chapter 3. To answer the first research question, in Chapter 3, we assess the relation between sentiment and stance in the stance dataset and evaluate the effectiveness of sentiment analysis methods in stance prediction. Then, we discuss the second research question in Chapter 4, where we provide an in-depth analysis of the possible online signals for stance detection on social media. To answer the third question, in Chapter 5, we develop a framework for stance obfuscation based on the most effective online features. To answer the fourth research question, we perform an empirical examination of probing the stance detection model to evaluate the presence of bots in conjunction with an online stance.

1.3 Contributions

This thesis makes four contributions to the literature, which can be categorized as the relation between sentiment and stance, stance modelling on social media, stance obfuscation and effect of online accounts on stance. The outcomes can be summarized as follows:

- **Examining stance modelling on social media and the relation between sentiment and stance.**
 - We survey stance modelling using textual and social interaction features that have been overlooked by previous stance studies.
 - We address the relation between the stance and sentiment. First, we explain the two tasks' definitions, where the stance differs from sentiment analysis as its target might not be explicitly mentioned or might not be the target of opinion. Then, we examine whether the supporting/opposing stances can be identified with positive/negative sentiments. We evaluate the effectiveness of the sentiment analysis method in inferring the stance across different domains. The finding of this study demonstrates that sentiment analysis is sub-optimal in inferring the stance and detecting support on social media.
- **Analysing factors for stance detection on social media.**
 - To evaluate the multi-modal representation of stance on social media, we analyze different online signals to evaluate the predictability of their stance towards multiple topics. We use a combination of content and network

interactions to demonstrate two main dimensions: linguistic and social interactions.

- Our finding shows the effectiveness of modelling the stance using network features along with the content of posts. Another finding shows the vulnerability of users on social media platforms, where their stance can be easily predicted. We demonstrate how users' stances can be easily detected even without them having to explicitly discuss the topic or posting online at all.
- **Stance obfuscation.**
 - We design a framework to help social media users preserve their privacy needs concerning the detection of stance online. This framework introduces two methods: data encapsulation and data removal. These two methods depend on the most predictive features in the stance detection model to generate divergents for the stance prediction results.
 - The findings demonstrate the effectiveness of data encapsulation in producing a divergent to the stance in comparison with data removal.
- **Assessing the effect of automated accounts (bots) on online stance.**
 - This study fosters more research to use stance detection by assessing the interplay between social bots and online stances. We provide a robust technique to analyse the interactions of bots with polarized stances by using a gold-standard stance dataset that covers various domains.
 - We conduct a comprehensive analysis on how users with specific stances are exposed to the bots' content through two types of interactions on social media: direct and indirect.
 - The findings show that a relation between social bot accounts and user stances does exist, but it is minimal when compared to other accounts.

1.4 Outline

Figure 1.1 illustrates the organization of this thesis along with the research questions. The rest of this thesis is organized as follows:

- Chapter 2. In this chapter, we examine the background pertaining to stance detection on social media. We conduct a thorough survey of different types

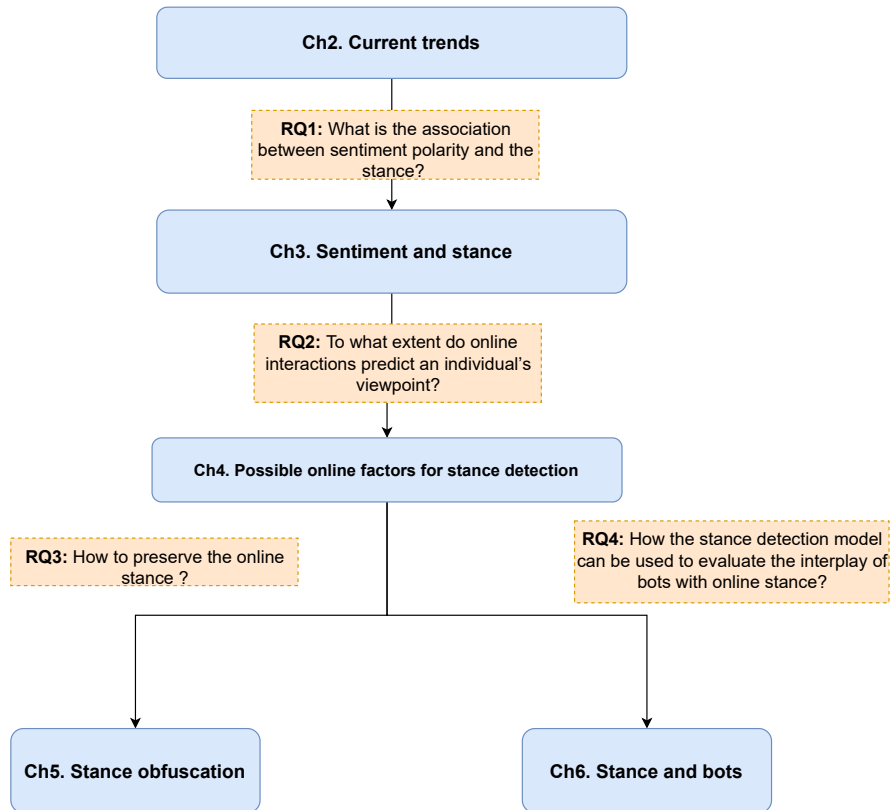


Figure 1.1: Thesis organization and the related research questions.

of stance detection and examine the current algorithms for stance detection on social media. This chapter begins with explaining various targets in stance detection as target-specific, multi-related-target and claim-based stance detection. Then, we review the current algorithms for stance modelling on social media, where we perform a comparative analysis of the current stance detection performance on different datasets and show the best stance modelling over the current well-known stance detection dataset, SemEval 2016.

- Chapter 3. This chapter distinguishes between sentiment and stance by providing an in-depth explanation of these two tasks. Then, we describe the construction of a new stance dataset to extend the current SemEval 2016 dataset by providing sentiment and stance for four new topics that cover different domains. This new dataset is constructed as context-dependent data, because the stance sometimes is only implied through context. Thus, we use conversational posts for the annotations instead of using a single post. Moreover, this study evaluates the

effectiveness of sentiment analysis methods for stance detection by using two datasets with different domains.

- Chapter 4. This chapter presents an analysis of the possible factors for stance detection on social media. We analyse various online signals to detect the stance towards politics, religion and other topics. This chapter compares multiple sets of online signals, including on-topic content, network interactions, user preferences and online network connections. We perform a thorough analysis of the most effective features in predicting online stance for five topics covering different domains.
- Chapter 5. In accordance with the findings in the literature, a stance can be detected without the users' having to express it explicitly. Thus, we propose a framework to help with preserving user stances on social media. This chapter demonstrates two methods of incorporating or removing some online signals to cipher the stance on social media.
- Chapter 6. The online stance can be inferred from a mixture of online features. By analysing the online social network interaction, we show how stance detection can be used to examine the interplay between bots and online stance. In particular, we analyse the bot account interactions with polarized stance and compare the overall impact with other real social media counts.
- Chapter 7. The chapter summarizes the main findings from previous chapters and proposes the future line of research.

1.5 Publications and research outcomes

- Aldayel, A., and Magdy, W. 2019. Your stance is exposed! analysing possible factors for stance detection on social media. Proc. ACM Hum.-Comput. Interact.3 (CSCW).
- Aldayel, A., and Magdy, W. 2019. Assessing sentiment of the expressed stance on social media. In Social Informatics (SocInfo), 277–286.
- Aldayel, A., Darwish, K., and Magdy, W. 2020 (Tutorial). "Detection and Characterization of Stance on Social Media", 14th International Conference on Web and Social Media (ICWSM).

- Aldayel, A., and Magdy, W. 2021. Stance detection on social media: State of the art and trends. *Information Processing and Management* 58(4):102597.
- Marcin Waniek, Abeer Aldayel, Talal Rahwan, Walid Magdy. "Obfuscating Opinions in social media", under-submission.
- Abeer AlDayel and Walid Magdy, "Characterizing the Role of Bots' in Polarized Stance on Social Media", under-submission.

Chapter 2

Background and related work

This chapter surveys stance detection on social media platforms. It maps out the terrain of existing research on stance detection and synthesizes its relation to existing theoretical orientations. Stance detecting on social media takes a way back, focusing on online debates in forums (Murakami and Raymond, 2010; Anand et al., 2011; Walker et al., 2012). This task has a rooted relation to argument mining, subjectivity analysis, and sentiment analysis (Ebrahimi, Dou, and Lowd, 2016). Online debate forums are considered a rich source of argumentative data, which has attracted many researchers to customize stance detection models for this platform. Stance detection for social media requires unique approaches to handle the noisy input containing informal and slang language.

Earlier work on stance detection focused on analyzing debates on online forums, which is distinguished from the more recent work that focused more on social media platforms, especially Twitter, since the former has a clear single context in comparison to social media platforms. In the online forums, the users debate in the form of a thread discussion where there is a flow of information usually focused on the topic Belkaroui, Faiz, and Elkhilifi (2014). In contrast, social media discussions on a given topic are more scattered, and sometimes they could be linked over a given hashtag Mohammad et al. (2016a).

2.1 Overview

The majority of work on stance detection has targeted the detection of the stance towards a given subject expressed in a given text. However, some works have studied the detection of the stance of users towards subjects without explicitly stating them, which

is usually referred to as stance prediction. Thus, the work on stance detection can be categorized into two main types: detecting expressed views vs predicting unexpressed views. In the first type, the objective is to classify a user's post and infer the current stance to be in favor or against a given subject Mohammad et al. (2016b).

In the later one, the prediction is carried out to infer the user's viewpoint on a given topic that the user did not discuss explicitly or towards an event that has not occur yet. This type of stance detection has proven its effectiveness in predicting the future attitudes in the aftermath of an event (Darwish et al., 2018; Magdy et al., 2016).

The work on stance detection can be also categorized based on the topic of the target of analysis, where it can be one specific target, multiple-related targets, or a claim in a news article. Most of the existing work designs stance detection classifiers to identify the user's stance towards one given specific topic. Sometimes the classifier is built to detect the stance towards multiple-related targets. This is the situation when a stance is detected towards two related entities that are typically opponents, such as detecting the stance towards Clinton and Trump simultaneously, since if the stance is in favor one target, it would be simply against the other (Sobhani, Inkpen, and Zhu, 2017). When the target is a claim in a news statement, the main objective is to identify if a given claim is supported by other posts. In this type of detection, the analyses is between two posts (source and reply) such as in RumourEval (Derczynski et al., 2017a,b) or between news header and other body of articles, which is mainly used as an initial step for fake news detection (Allcott and Gentzkow, 2017)

Therefore, this chapter provides a survey study to identify the current trends in stance detection on social media. The chapter is organized as follows: Section 2.2 provides theoretical definition of stance. Sections 2.3 and 2.4 summaries the literature on stance prediction and categorizes this work according to the type of detected stance (expressed vs unexpressed) and the target of the stance in the text. Sections 2.5 explains the different stance modeling and machine learning approaches for stance classification. Section 2.6 lists several applications for stance detection, such as social media analysis and fake-news detection. Section 2.7 lists the current available resources for stance detection tasks. Finally, the current research trends on stance classification are discussed in section 2.8 while highlighting the gaps and suggesting the potential future work required in this area.

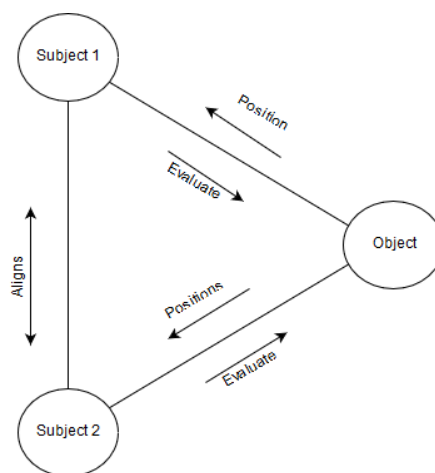


Figure 2.1: The stance triangle, adapted from (Du Bois, 2007)

2.2 Stance definition

Biber and Finegan 1988 define stance as the expression of the speaker's attitude, standpoint and judgment toward a proposition (Biber and Finegan, 1988).

Du Bois (2007) argues that stance-taking (i.e. a person taking a polarised stance towards a given topic) is a subjective and inter-subjective phenomenon in which stance-taking process is affected by personal opinion and non-personal factors such as cultural norms. Stance taking is a complex process relates to different personal, cultural, and social aspects. For instance, political stance taking depends on experiential behavior as stated by (McKendrick and Webb, 2014).

The process of detecting stance of a given person on social media is still in its infancy as it is not yet clear what the role of language and social interaction plays in inferring the user's stance.

Stance detection has a strong history in sociolinguistic, where the main concern is to study the writer's viewpoint through their text. Stance detection aims to infer the embedded viewpoint from the writer's text by linking the stance to three factors, namely: linguistic acts, social interactions, and individual identity. Using linguistic features in the stance detection is usually associated with attributes such as adjectives, adverbs and lexical items (Jaffe, 2009).

It has been argued that stance taking usually depends on experiential behavior, which is based on previous knowledge about the object of evaluation (McKendrick and Webb, 2014). This strengthen the stance detection as a major component in various analytical studies.

Stance detection on social media concerns with an individual's views towards an

Table 2.1: Examples of stance from SemEval2016 stance dataset

#	Tweet	Target	Stance
1	Great comments by @MittRomney about pathological liar #HillaryClinton .	Hillary Clinton	Against
2	Life is sacred on all levels. Abortion does not compute with my philosophy. (Red on #OITNB)	Legalization of Abortion	Against
3	Also a policy on removing feminists. In their entirety. Thanks for your help with this VoteUKIP	Feminist Movement	Against
4	If it's been 80 days since you've done an interview, can we assume you're not really running for president?	Hillary Clinton	Neither
5	Everyone is able to believe in whatever they want. Freedom	Atheism	Favor
6	Reminds me of Nottingham in the early 1960's #ActOnClimate	Climate Change is real threat	Favor

object of evaluation by using various aspects related to the user's post and personality traits.

As stated in Due Bois's stance triangle shown in Fig 2.1, the process of taking a stance is based on three factors: 1) Evaluating objects; 2) Positioning subject (self); and 3) Aligning with other subjects (other social actor). For instance, "I am with the new legalization on Climate Change" has a subject "self" indicated by proposition "I" and the "with" indicates the favor position towards the object "Climate Change". The complexity of stance interpretation in social media is stemmed from multiple elements that affect stance-taking process. In social media, identifying the stance subject (the self) is mostly straightforward as each post is linked to the user.

Furthermore, from sociolinguistic perspective (Jaffe, 2009), it has been argued that there is no complete neutral stance as people tend to position themselves through their texts as in favor or against the object of evaluation. This casts a further complexity in identifying the stance of the social actors, since the stance is not usually transparent in the text, but sometimes needs to be inferred implicitly from a combination of interaction and historical context.

2.3 Stance detection tasks

From literature, stance detection has multiple forms, which can be categorised: 1) according to the target type, whether it is a single target, multi-related-targets, or claim-based targets; or 2) according to the classification task itself, where it is a detection of an existing stance or prediction of a future stance. In the following, we initially make a distinction between the level on which stance prediction is applied, then we discuss the different tasks of stance detection.

2.3.1 Stance detection according to target

Stance detection is the task of inferring the viewpoint either in support or against a given topic or entity. Therefore, stance detection needs the presence of a defined target to detect the stance towards it. In the literature, stance detection can be be categorised according to the type of the target of evaluation into three categorizes, namely: single defined target, multi-related targets, and claim-based. In the following, a further explanation on each type is provided.

Target-specific stance detection

For stance detection, a clear target G needs to be defined in advance to assess the overall attitude towards this target.

The basic form of Stance detection on social media can be formulated by using the attributes of the social actor. Thus, in this form of stance detection, the main two inputs are 1) text T or user U , and 2) given target G as illustrated in equation 2.1.

$$\text{Stance}(T, U|G) = \{Favor, Against, None\} \quad (2.1)$$

There are three possible variations of this definition which may include as input either the text, the social actor or both for stance detection. The dependent factor in stance detection task is the *target* of analysis. One of the common practices in stance detection on social media is to infer the stance based on the raw text only, which maps the stance detection problem to a form of a textual entailment task (Lin et al., 2006; Mohammad et al., 2016b). From a more relaxed definition, a text T entails a stance to a target G , ($T \rightarrow \text{stance to } G$), if the stance to a target can be inferred from the given text. In social media analysis, this formulation has been further extended to include the social actor which could be represented by a set of online behavioral features to infer

the stance. Instead of the dependency on the raw text, the other representation of stance detection includes the social actor U as a salient factor to infer the stance. Reflecting on the origin of stance taking using the stance-triangle as defined by Due Bois's stance triangle shown in Fig 2.1 (Du Bois, 2007), where the social actor (subject) is the main element of this process. The structure of social media extends the possibility to represent the social actor by using a variety of network features such as profile information (age, description,..etc) (Magdy et al., 2016; Darwish, Magdy, and Zanouada, 2017a).

For instance, in table 2.1, example 1, shows an *against* stance towards a given target *Hillary*. As Equation 1 shows the main input components as the text (*Tweet*), along with target (G). Additionally, including social actor attributes helps in identifying the holder views. This kind of modeling has been heavily used in various studies where the homophily of the network is used to identify the users views (Garimella and others, 2018; Borge-Holthoefer et al., 2015). The target-specific stance detection is the basic practice for various stance studies, even for benchmark datasets such as SemEval stance 2016, which is covering multiple topics, most of the published work on this dataset trained a separate model for each topic (target) separately (Mohammad et al., 2016b; Aldayel and Magdy, 2019b; Siddiqua, Chy, and Aono, 2018). However, there have been recent trials to apply the transfer learning along with other unconstrained supervised learning on different targets as will be discussed in more details in section 2.5.2.

One of the well known stance detection benchmark dataset is the **SemEval stance 2016**. This dataset contains tweets cover different domains. The tweets collected based on a predefined hashtags and keywords that are related to the topic. The task has been defined as textual entailment task, where for a given tweet the aim is to infer the stance towards a target. Then each tweet is annotated with the stance as in-favor, against or neither. This dataset has been introduced in SemEval 2016, which includes two sub-tasks namely, sub-task A (supervised learning) and sub-task B (weak supervision stance detection). In task A, the dataset contains about 4000 tweets labeled with stance and sentiment towards five topics: Hilary Clinton (HC), Feminist Movement (FM), Atheism (A), Legislation of Abortion (LA) and Climate change (CC). About 19 teams participated to detect the stance in about 4000 tweets. Each topic has been annotated with sentiment and stance labels. Overall, the against stance constitutes around 49.47 % of the dataset, with 62.30 % of tweets are expressing negative stance. We further explain the property of this dataset in Chapter 3, section 3.4.1.

Multi-related-targets stance detection

In multi-target stance detection, the goal is to jointly learn the social media user orientation towards two or more targets for a single topic Sobhani, Inkpen, and Zhu (2017). The main assumption behind this kind of stance detection is that when a person gives his stance for one target this provides information about his stance towards the other related targets.

$$\text{Stance}(T|U, G_n) = \{(FavourG_1, AgainstG_{n+1}), (FavourofG_{n+1}, AgainstG_1)\} \quad (2.2)$$

For instance, a tweet could express a stance toward multiple US presidential candidates at the same time; thus, for example, when a user expresses their in-favor stance towards Trump, it implies an against stance toward his opponent Darwish, Magdy, and Zanoouda (2017b); Lai et al. (2016). In Sobhani, Inkpen, and Zhu (2017) work, the first multi-target stance detection data-set is presented, containing 4,455 tweets related to the 2016 US elections. In order to detect the subjectivity toward two targets, Sobhani, Inkpen, and Zhu (2017) used an attention-based bidirectional recurrent neural network (RNN) to jointly learn the stances toward Clinton and Trump. The notion of multi-target stance detection has been usually used to analyze the relation between two political candidates by using domain knowledge about these targets to improve the classification performance Lai et al. (2016). Following the same approach, the study by Darwish, Magdy, and Zanoouda (2017b) constructed a dataset with 3,450 tweets annotated with stance labels for the two US 2016 election candidates (Trump and Clinton) at the same time. Furthermore, the work of Wei, Lin, and Mao (2018) proposed a memory-based algorithm focusing on jointly modeling multiple targets at the same time. Their memory-based model provides the current state-of-the-art result so far on the multi-target benchmark dataset.

Claim-based stance detection

In claim-based, also known as open-domain stance detection, the target of the analysis is not an explicit entity as the ones discussed earlier; however, it is a claim in a piece of news. The first stage to detect the stance is to identify the main target claim from the sequence of conversation or given text. The main input to the claim based stance detection model is the claim (C) which could be the rumour's post or article headline based. In the fake news task, the claim tends to be the article headline and the text is the article body. On the other hand for the rumour's veracity task, the main input to be

evaluated is the rumour's post and the text is the reply to the rumours. The prediction label sets tend to take the form of confirming the claim or denying it. For instance, for a given claim "These are not timid colours; soldiers back guarding Tomb of Unknown Soldier after today's shoot-ing StandforCanada". The given reply post (T) "Apparently a hoax. Best to take Tweet down".

$$\text{Stance}(T,C) = \{\text{Confirming}, \text{Denying}, \text{Observing}\} \quad (2.3)$$

Claim-based stance detection is considered a suitable method to analyze the veracity of the news. For that reason, claim-based stance detection has been heavily used for rumor resolution studies Hamidian and Diab (2015); Aker, Derczynski, and Bontcheva (2017); Zubiaga et al. (2018); Derczynski et al. (2017a). In study by (Hamidian and Diab, 2015), they used a supervised learning model along with a new set of features called "pragmatic features" which contains: named entity, event, sentiment and emoticons. Interestingly, (Aker, Derczynski, and Bontcheva, 2017) concluded that problem-specific features engineering outperformed other state-of-the-art systems in rumour identification tasks. Their model, which used a random forest outperformed the advanced LSTM-based sequential model in SemEval 2017 task 8 proposed by (Kochkina, Liakata, and Augenstein, 2017). Within the same line, the study of (Aker, Derczynski, and Bontcheva, 2017) used the same feature engineering approach proposed by (Elfardy and Diab, 2016), in which lexical and semantic features used to help with identifying the stance of a Twitter user. More recently, for conversation-based tasks, the study by (Zubiaga et al., 2018) showed that using LSTM can outperform other sequential classifiers and feature-based models.

In another study by (Li, Zhang, and Si, 2019) they used multi-task learning with stance detection layer to classify the stance of a given tweet as supporting/ denying a given claim.

2.3.2 Stance detection vs stance prediction

Another possible categorisation to the stance detection task can be framed according to the status of the stance to be modeled, either being 1) an existing stance expressed in text; or 2) an unexpressed stance that might have not occurred yet. The first type of work is generally referred to as stance classification/detection, and it represents the majority of work in the existing literature. The other type of work is usually concerned with detecting the stance prior to an event, and thus is referred to as stance prediction.

Appendix A lists the various approaches that have been used for stance detection on social media.

Stance detection for expressed views

The underlying method for this type is based on using a predefined set of keywords concerning the target of analysis. For instance, the SemEval stance dataset (Mohammad et al., 2016b) created three sets of hashtags for each event or entity to collect tweets concerning three stances. For example, for the topic "Hillary Clinton," the dataset used favor hashtags, such as #Hillary4President, against hashtags, such as #HillNo, and ambiguous hashtags. Ambiguous hashtags contain the target without a direct indication of the stance in the hashtag (e.g., #Hillary2016). In a study done by (Darwish, Magdy, and Zanouda, 2017b) to predict public opinion and what went viral during the 2016 US elections, tweets that were filtered out by using a set of 38 keywords related to the US elections for streaming relevant tweets to this event were used. Another study employed tweets generated by users to infer the opinion toward the independence of Catalonia and collected tweets by using two keywords related to the hashtags #Independencia and #27S (Taulé et al., 2018a). Similarly, to study mass shooting events, a study by (Demszky et al., 2019) made use of Gun Violence Archive and collected list of events between 2015 and 2018 to define a set of keywords that could be used to retrieve the tweets.

In these studies, after collecting the event-related data, it gets annotated using predefined guidelines for labelling the stance toward the given target (Mohammad et al., 2016b).

Stance prediction of unexpressed views

Stance prediction aims to infer the stances of social media users with no explicit expression of these stances online. It is also used to predict stances on events that have not occurred yet.

In stance prediction, most studies tend to predict the users' stances on two levels, that is, micro and macro. At the micro level, stance prediction means the estimation of an individual user's standpoint toward a target or event in advance (pre-event), indicating the likelihood that the user will be in favour of or against the target event. This methodology is similar to recommending new items based on a user's history of purchases. At the macro level, the public opinion toward an event is inferred, and

the research tend to address this level of prediction as an aggregation of micro predictions Qiu et al. (2015). Similarly, Dong et al. (2017) designed a micro-level stance detection model based on users' previous posts along with a logistic regression to predict stances for new users that were not included in the dataset. Their model produced word distributions for the against/support stances on five topics: Bowe Bergdahl, Gaza Israel, Immigrant, Hobby Lobby, and MH1, by using rules derived from user's interactions. Another study by Gu et al. (2014) used matrix factorization to predict micro-level stances based on the voting patterns of the users and the topic model. The study by Gu et al. (2017) predicted ideology using heterogeneous links between the users.

The work by Darwish, Magdy, and Zanzoua (2017a) investigated the best social media features to predict user attitudes toward new events. This work differed from the SemEval-2016 task 6 (a target-specific task) in that the stance of the users was predicted based on each user's historical collection of tweets, instead of a target-relevant single tweet. Rather than focusing on the tweet content (textual features), Darwish, Magdy, and Zanzoua (2017a) used the similarity between users and inferred latent group beliefs as features in their model, and further employed a bipartite graph along with graph reinforcement. A different method was proposed by Gottipati et al. (2013) to predict an individual's unexpressed stance by implementing a collaborative filtering approach and constructing a user ideology matrix that represents a user's ideological stance for an event.

In macro-level stance prediction, few studies have addressed stance prediction. Most of these studies were customized to analyze specific cases, such as Islamophobia Magdy et al. (2016); Darwish et al. (2018). Besides analyzing events, these studies carried out a further complementary study for the prediction of user's stances in the aftermath of an event (Paris terrorist attacks in 2015), based on previous tweets and user interaction. In debate forums, Qiu et al. (2015) have proposed a micro-level stance prediction model based on user behaviour toward new events in which they have not participated. In this study, this kind of user was referred to as a cold-start user, which is a well-known term commonly used in recommendation systems. In addition, they introduced a macro-level stance prediction model as an aggregation of each user's stance (i.e., at the micro-level).

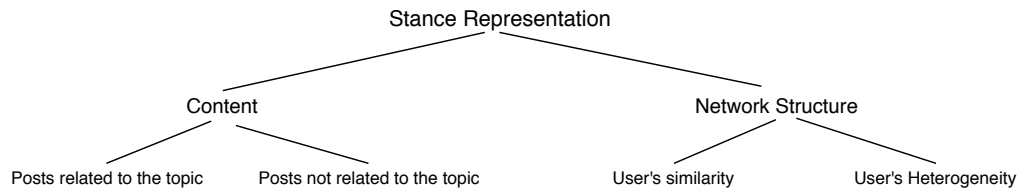


Figure 2.2: Stance representation in social media

2.4 Stance modeling on social media

Stance on social media has been modeled using various online signals, which have been used as features for training the stance detection models. Those signals can be categorised into two main categories: 1) content signals, such as the text of the tweets; and 2) network signals, such as users' connections and interactions on their social networks. Figure 2.2 summarises the main features used in each category to model the stance on social media. In the following we describe each of these two categories of signals and how they have been used for detecting stance. Finally, we present a comparison to their effectiveness for detecting stance on multiple datasets.

2.4.1 Content features

This section discusses the stance representation focusing on the textual features derived from the user's content online. As illustrated in Figure 2.2, the content can be collected based on the topic of the analysis. In this kind of feature representation, the data is collected based on range of keywords reflecting the topic. Another type of representation concerns with collecting content that has no direct relation to the topic. The main objective in this case is to model the stance based on the user's behavioural data rather than topic level stance detection.

Furthermore, the content-features can be categorized into two general types: linguistics features and user's vocabulary. The first type of features concerns with the text linguistic features that helps in inferring the stance. The other type concerns with modeling user's stance based on the user's choice of vocabulary.

Linguistics features

The majority of the work on stance detection focused on utilizing the linguistic elements that capture the social media user's stance. This massive dependency on linguistic cues is due to defining stance detection as textual entailment task, where the

task is to detect a stance in a given piece of text (e.g. tweet Mohammad et al. (2016b)). In the literature, the stance detection work that concerns with using textual cues to detect stances includes: textual features, sentiment polarity, and latent semantics.

For instance, using textual features such as n-gram modeling of the text has been heavily investigated in the literature (Anand et al., 2011; Mohammad, Sobhani, and Kiritchenko, 2017; Sobhani, 2017). Using the n-gram modeling of the text shows promising results. In the SemEval 2016 stance detection task, using word and char n-gram modeling managed to get the best f-score among the other participating systems (Mohammad et al., 2016b). Another textual feature that have been used to infer the stance is the sentiment polarity of the post (Mohammad, 2016; Elfardy and Diab, 2016; Elfardy, 2007). In general, the use of sentiment as feature was not sufficient in predicting the stance as studies concluded (Mohammad, Sobhani, and Kiritchenko, 2017; Elfardy, 2007). Another kind of features is the latent semantics features which aims to reduce the dimension of a given input such as mapping the sentences to pre-defined set of topics (topic modeling) (Elfardy and Diab, 2016). Topic modeling has been applied by different studies (Elfardy and Diab, 2016). For instance (Elfardy and Diab, 2016) used Textual Weighted Textual Matrix Factorization (WTMF) and frame-semantic parser to model a given tweet has been used as feature by (Patra, Das, and Bandyopadhyay, 2016) to map unigrams to topic sphere. Another work used a simple bag of topic to model targets words.

User's vocabulary choice

There is a considerable number of studies that represent the stance based on the user's vocabulary. The hypothesis behind using this kind of user modeling is that individuals with the same stance tend to use the same vocabulary choice to express their point of view Darwish et al. (2020a).

The focus of these studies is mainly to disentangle the topic from the viewpoint where the vocabulary is not only linked to the topic but to the individual attitude and characteristics (Beigman Klebanov, Beigman, and Diermeier, 2010). For instance, people with against stance on abortion tend to use vocabulary such as pro-life to express their opposing stance. The work of (Beigman Klebanov, Beigman, and Diermeier, 2010) built a user stance detection model based on users vocabulary choice using a generative model and discriminative model using Naive Bayes and SVM classifiers, respectively. Another work by (Dong et al., 2017) proposed a word generative model based on the user interaction to build a set of word representation for each stance

on a topic. In addition, the work of (Benton and Dredze, 2018) follows the same direction where the users' interactions help in shaping the set of vocabularies used to identify the stance. Another study by Zhu, He, and Zhou (2019) used a hierarchical topic for opinion discovery based on the author of the text vocabulary. Lately the work on vocabulary choice has been linked to user behaviour in the social media to further improve the stance prediction model. The work of (Dong et al., 2017) used a generative model based on the user's keyword choice for each standpoint along with users interactions as a regularization to predict the stance. In addition (Li, Porco, and Goldwasser, 2018), introduced a users interaction and post embedding by using an embedding vector for the Pro-stance and another vector for the Con-stance.

2.4.2 Network features

Social media provides a special kind of social data due to the structure of these platforms, where users can be characterised and/or analysed based on their social connections and interactions.

Many existing work used network features to gauge the similarity between the users.

The network features that have been used to learn users representations in social media can be grouped under two categories: user behavioral data (Thonet et al., 2017; Darwish, Magdy, and Zanoluda, 2017a; Darwish et al., 2020b), and user meta-data attributes (Pennacchiotti and Popescu, 2011). Using users behavioral data to identify the stance is motivated by the notion of homophily, which is based on the social phenomenon "individuals associate with similar ones" Bessi et al. (2016). In social media, the similarity between users considered a core property that helps in inferring stances.

The interaction elements have been used to define the similarity between the users. One of these elements that has been extensively used to infer Twitter user's stance is the retweet (Borge-Holthoefer et al., 2015; Darwish et al., 2018; Weber, Garimella, and Batayneh, 2013; Rajadesingan and Liu, 2014). Another element that has been heavily investigated is hashtags, this element has been used in the literature to infer similarity between users in order to predict the stance (Darwish, Magdy, and Zanoluda, 2017a; Dey et al., 2017). The work of (Dey et al., 2017) used soft cosine similarity to gauge the similarity between the users who post on the same hashtags. The work of (Darwish, Magdy, and Zanoluda, 2017a) used graph reinforcement to calculate the similarity between the users who post on the same hashtags.

In a recent study by (Aldayel and Magdy, 2019b) they defined three types of network features to model the stance on social media. Those network features are: 1) interaction network, 2) preferences network and 3) connection network. The interaction network represents the users direct interactions with other users in the sense of retweet, mention and reply. This type of network provides the best performance score of stance detection model in comparison with the other two networks. The preference network is the network of others users that post or are mentioned in the tweets the user likes. This network allows detecting stance for users who might have limited posting or interaction behaviour online. Finally, the connection network includes the friends and followers of the users. The three types of networks provide the best performance in comparison with content features 2.2.

Another study by (Fraisier et al., 2018) introduced a multi-layer graph model to represent profiles from different platforms and to extract communities. By doing this, the model allows stance to diffuse from the set of known profiles and to predict profiles' stance by using pairwise similarities between users' profiles.

The use of network features has shown its benefit in detecting social media user stance and future attitudes in the aftermath of an event (stance prediction). For instance, in study done by (Darwish et al., 2018) they used the user's similarity components as features. In their work, the similarity was calculated based on the interaction elements of a given tweet. These interaction elements are: mentions, re-tweets, and replies; Website links (URLs) and hashtags used by users in their tweets. Similarly, the work of Thonet et al. (2017) used the retweet and reply to define the user's social network.

Another line of study, use heterophily to measure the dissimilarity between users to model the stance (Trabelsi and Zaïane, 2018). For instance the study of (Trabelsi and Zaïane, 2018) used the tendency of a user to reply to the opposed viewpoint. The study used a rebuttal variable to model the users interactions and denote if the reply attacks the previous author parent post. The value of the rebuttal depends on the degree of opposition between the viewpoint of parent post and the parent tweet.

2.4.3 Comparative analysis of stance modeling

Table 2.2 shows a comparative analysis of the stance detection performance using two types of modeling, content vs network. In addition, it reports the results of studies that used a combination of both. It can be noticed that using network features to model

the stance outperforms the content modeling of the stance. The Textual modeling has the lowest performance as this type of modeling depends on textual entailment task, while the network features provide the highest performance in comparison with the later. This can be as a reason of the fact that using network features puts into consideration the bigger picture of modeling the user's attitude using the users' interactions and similarity. Using Network features overcomes the limitations poses by textual entailment modeling of stance. As demonstrated by the (Aldayel and Magdy, 2019b)'s study, where they used stance SemEval dataset to model stance on social media using the model with best reporting results on the stance dataset to compare the performance of stance when using content and network features. Their study provides a thorough comparative analysis of the stance overall performance when using textual and network features. The use of network features provides the best performance in compression in previous studies, where they used textual modeling along with transfer learning methods as reported by (Mohammad et al., 2016b). Furthermore, in study by (Lahoti, Garimella, and Gionis, 2018) they show that the network features outperform the textual modeling of stance when using non-negative matrix factorization for stance detection. The study of (Darwish, Magdy, and Zanoouda, 2017a) confirms the same conclusion where the use of network features outperforms the textual representation using two datasets (Islam and Island data-sets). As the study by (Magdy et al., 2016) shows that network modeling of stance outperforms the use of textual modeling for the expressed and non-expressed stance on social media.

2.5 Stance detection algorithms

In this section, the main machine learning (ML) algorithms used for stance detection is discussed. According to the work in literature, the ML algorithms used for stance detection on social media can be divided into three main approaches: 1) supervised learning; 2) Weakly-supervised and transfer learning; and 3) unsupervised stance detection. In the following, each of these approaches are discussed in more detail.

2.5.1 Supervised learning

This is the basic and most common approach for most of the work on stance detection (Zhang et al., 2019; Lai et al., 2020a; Walker et al., 2012; Krejzl and Steinberger, 2016; Igarashi et al., 2016; Gottipati et al., 2013). With this approach, a stance

Table 2.2: A comparison of stance detection models using network , content and both as features.

Dataset	NW	Content	Both	Model
Before Paris attacks (Magdy et al., 2016)	85.0	82.0	84.0	SVM
Islam (Darwish, Magdy, and Zanouada, 2017a)	84.0	76.0	-	SVM
Islands (Darwish, Magdy, and Zanouada, 2017a)	79.0	71.0	-	SVM
Gun control, abortion and Obama care (Lahoti, Garimella, and Gionis, 2018)	81.9	60.6	82.0	NNMF
SemEval stance (Aldayel and Magdy, 2019b)	71.6	69.8	72.5	SVM
SemEval stance (Lynn et al., 2019)	61.8	62.8	65.9	SVM
LIAC (Dong et al., 2017)	67.8	54.1	-	Generative Model

dataset is annotated using a predefined set of labels, usually two (pro/against) or three (pro/against/none) labels. For instance, the SemEval 2016 stance dataset uses three labels: 'In-Favor', 'Against' and 'None' (Mohammad et al., 2016b) for a set of five different topics. Many studies have been published on this dataset used different supervised ML algorithms such as classical algorithms, such as (Naive Bayes) NB, SVM, and decision trees; and deep learning algorithms, such as RNNs and LSTMs, to detect the stance on this labeled dataset.

For example, the work of (Mohammad, Sobhani, and Kiritchenko, 2017) used a SVM model with linguistic and sentiment features to predict the stance. Their study showed that the use of content features only (n-gram) provides an F1 score equal to (69.0%) which surpassed the use of sentiment feature with about 66.8% F1 score. Another work by (Walker et al., 2012) used an labeled data toward 14 topics to train a NB model to predict the stance. The work of (Elfardy and Diab, 2016) used SVM model and lexical and semantic features to classify the stance in SemEval stance which has F1 score equal to 63.6%. Another study by (Wojatzki and Zesch, 2016) used stacked classifier and syntactic features to classify the stance in SemEval stance dataset. Their models show a minuscule improvement on the overall stance detection performance with about (62%) F1 score. The work of (Siddiqua, Chy, and Aono, 2019a) proposed neural ensemble model using bidirectional LSTM on SemEval stance dataset along with a fast text embedding layer. Their model shows an improvement on the overall

F score to 72.1%. A recent work by (Li and Caragea, 2019) used bidirectional gated recurrent unit to build a multitask learning model that leverages the sentiment of a tweet to detect the stance. This model shows an improvement of stance detection model overall F score of Semeval stance dataset to reach a score equal to 72.3%. A detailed comparison of each study performance on Semeval stance dataset is reported later in section 2.5.4.

2.5.2 Unconstrained supervised learning

To address the scarcity of the labelled data for each target in the stance detection task, some studies in this field attempted to incorporate unconstrained supervised methods, including transfer learning, weak-supervision, and distant supervision methods for stance detection.

Several studies applied transfer learning techniques to enrich the representation of the target entities in the dataset and enhance the overall performance of stance detection (Dias and Becker, 2016; Augenstein, Vlachos, and Bontcheva, 2016). In transfer learning, the knowledge that an algorithm has learned from one task is applied to a separate task, such that in the task on stance detection, transfer learning is applied across different targets. One of the well known stance detection datasets that motivated the work on transfer learning is the SemEval stance (Task B) (Mohammad et al., 2016b). This dataset contains 707 labeled tweets and 78,000 unlabelled tweets related to Trump and provides a good source for research, to be explored through various transfer-learning algorithms. For example, in the work of (Augenstein, Vlachos, and Bontcheva, 2016), the SemEval stance (Task B) was used to train a logistic regression (LR) model along with a bag-of-words auto-encoder along with Hillary-labelled data to detect the stance. Another study by (Dias and Becker, 2016) used Hillary Clinton's labelled data along with Trump's unlabelled data to develop a rule-based algorithm to help in detecting the stance for the SemEval stance's Task B. In the work of (Wei et al., 2016), a CNN model was used with Google news embedding on the SemEval stance's Task B. For Subtask B, their model trained on two class datasets and further, by using a modified softmax layer, the classification of three classes with a voting scheme was performed. With these configurations, their model ranked among the top three models with an overall F-measure of 56.28 for Task B. Moreover, the study by (Zarrella and Marsh, 2016) implemented distant supervision to predict the stance using SemEval (Task A) dataset by using a recurrent neural network (RNN) and pre-training the recurrent layer using

a distant supervision with a hashtag prediction as auxiliary task. They used unlabelled datasets collected from 197 hash-tags with relevance to the topic. This method was effective to enhance the performance of stance detection trained on limited data by transferring features from other systems trained on large unlabeled datasets.

Many studies have incorporated the concept of transfer learning using new datasets other than the SemEval stance (Task B). A recent study by (Hanawa et al., 2019) used Wikipedia articles to build knowledge extraction for each topic on a dataset that contained seven topics from Wikipedia. Another work by (Ferreira and Vlachos, 2019) used three datasets, namely blogs, US elections, and Moral Foundations Twitter, and designed algorithms that incorporated label dependencies between them to train the stance detection model. Recently study by (Ghosh et al., 2019) shows that the use of BERT model for stance detection is providing the most effective performance on two stance datasets: the SemEval 2016 stance dataset (Mohammad et al., 2016b) and set of online news articles (Sen et al., 2018). The study by (Zhang et al., 2020) used SemEval stance task A and B along with new dataset '*Trade Policy*' to construct eight cross-target stance detection sub tasks based on splitting the the task into two groups: Women's Right (Feminist Movement, Legalization of Abortion) and American Politics (Hilary Clinton, Donald trump, Trade Policy). In their study, they used knowledge-aware memory components to incorporate the external knowledge into BiLSTM.

Recently, (Giorgioni et al., 2020) introduced a transformer-based architecture along with data augmentation and ranked first in the Italian tweets (Sardistance) task within the context of EVALITA 2020 (Cignarella et al., 2020). In their model they trained a specific UmBERTo based sentence classifier on three auxiliary datasets from three tasks, sentiment, irony and hate-speech, and the resulting labels augmented as a new sentence in the SardiStance dataset. Then, they augment the training dataset by labeling additional tweets using distant supervision over a specific set of hashtags. To predict the stance towards the 2020 presidential candidates, Joe Biden and Donald Trump, a recent study by (Kawintiranon and Singh, 2021), fine tuned BERT with an unlabeled in-domain dataset related to US 2020 elections. They introduce a knowledge enhanced masked language modeling by training the transformer encoder with masked language modeling in many BERT-based models. In their method, they used attention-based language model to pay attention towards distinctive words of each stance.

2.5.3 Unsupervised learning

Recently, the attention has been devoted toward building an unsupervised stance detection model. In this kind of studies, they mostly used clustering techniques with focus on the user and topic representation on the social media platform (Darwish et al., 2020a; Joshi, Bhattacharyya, and Carman, 2016; Trabelsi and Zaïane, 2018). The work of (Trabelsi and Zaïane, 2018) proposed unsupervised models using a clustering model at the author and topic levels. In their study, they used six topics collected from two online debate forums 4Forums and Create-Debate. Their clustering model leverages the content and interaction network of the users (retweets and replies).

A recent study by (Darwish et al., 2020a) used clustering technique to create an initial set of stance partition for the annotation. In their work, they used unlabeled tweets related to three topics: Kavanaugh, Trump, and Erdogan. Their findings show that using retweets as feature provides the best performance score when implementing clustering algorithm (DBSCAN) which surpass the supervised method when using fast-text and SVM model. Their finding is considered a large motivation for using unsupervised methods for stance classification in the future.

2.5.4 Best performing algorithm

Since the SemEval stance dataset was the most used dataset for bench-marking the performance of stance detection by using multiple ML approaches, this dataset is used in this section to compare different ML approaches discussed in the above sections. Table 2.3 shows the best performing models based on the type of algorithm. As can be expected, the transfer learning models have the lower performance score compared to supervised learning models. However, the performance by (Li and Caragea, 2019) shows promising performance with an F-score of 0.653. However, some other trials, such as the work of (Ebner, Wang, and Van Durme, 2019), which uses the Deep averaging network (DAN) with the Glove word embedding, achieves an average F1 score is around 0.3%, which is close to random performance.

As can be noticed clearly from Table 2.3, the supervised algorithms are more effective for stance detection, where they achieve higher F-scores on the SemEval dataset than transfer-learning approaches. This can be seen in the models that incorporated Network features to detect the stance. It is interesting to see that simpler machine learning models, such as SVM are more effective than deep learning models. In addition, these models achieve even better performance when incorporating network fea-

Table 2.3: Comparing the Stance detection models on SemEval stance dataset.

Algorithm	Model	Features	F1	Study
Supervised learning	SVM	NW	71.56	(Aldayel and Magdy, 2019b)
	SVM	NW+content	72.5	(Aldayel and Magdy, 2019b)
	RNN	NW	61.8	(Lynn et al., 2019)
	RNN	NW +Content	65.9	(Lynn et al., 2019)
	LSTM	Content	72.1	(Siddiqua, Chy, and Aono, 2019a)
	Bi-GRU	Content	72.33	(Li and Caragea, 2019)
	HAN	Content	69.8	(Sun et al., 2018)
	SVM	Content	70.0	(Siddiqua, Chy, and Aono, 2018)
Transfer learning	Bi-GRU	Noisy labeling + topic modeling	60.8	(Wei, Mao, and Chen, 2019)
	Bi-LSTM	Sentiment lexicon	65.3	(Li and Caragea, 2019)
	DAN	words embedding	35.2	(Ebner, Wang, and Van Durme, 2019)
	DAN	Glove	30.2	(Ebner, Wang, and Van Durme, 2019)

tures along with content, where using both with a simple linear SVM is more effective than using word-embedding with RNNs and LSTMs.

There is no unsupervised algorithms that have been applied to the same SemEval dataset till date. The only reported results as discussed earlier are the ones by (Darwish et al., 2020b), where the model is applied on different dataset (Trump, Kavanaugh and Erdogan datasets). The unsupervised model of this study used the network features and outperformed the model with the use of fast text word embedding with clustering purity equal to 98%. Overall, it can be noticed that the best performing models are the one framed in user and social context with use of user network features in the stance detection task.

2.6 Stance detection applications

Stance detection has been mainly used to identify the attitude towards an entity of analysis to allow measuring public opinion towards an event or entity. However, there are other applications for stance detection that are discussed in this section.

2.6.1 Analytical studies

Using stance detection has proven its benefit as social sensing technique to measure public support related to social, religion and political topics.

As examples of using stance detection for analysing political topics are studies on analysing public reaction towards Brexit using Twitter data (Lai et al., 2020b; Simaki et al., 2020; Grcar et al., 2017; Simaki, Paradis, and Kerren, 2017). Furthermore, most of the stance detection studies used US 2016 election data to analyze the stance toward two candidates, Hillary Clinton and Trump (Lai et al., 2016; Darwish, Magdy, and Zanouda, 2017b). The work of (Lai et al., 2018) analysed the stance towards the political debates in Twitter about the Italian constitutional referendum held on December 2016. Another work by (Taulé et al., 2017) studied the public stance toward Catalan Independence. A more recent study by (Lai et al., 2020b) used a multilingual dataset to study political debate on social media. They collected entities and events related to politics in five languages: English, French, Italian, Spanish and Catalan.

Another line of studies used stance detection to analyse the public viewpoint towards social aspects. The public opinion towards immigration has been recently studied (Gualda and Rebollo, 2016; Bartlett and Norrie, 2015). The work of (Gualda and Rebollo, 2016) studied the attitude towards the refugees using twitter. They collected tweets using the keyword "refugees" in different languages. They used the sentiment of the discourse to analyze the public stance towards the refugees. Another work by (Bartlett and Norrie, 2015) studied the immigration in the United Kingdom by annotating the tweets related to immigration with sentiment polarity. They used the negative polarity as an indication of the against stance and the positive sentiment as an indication of the support viewpoint. They found that about 5% of the users were against immigration, while 23% were supporting the immigration. It worth mentioning that the last two mentioned studies are seen suboptimal in measuring stance due to relying heavily on sentiment of the text, which will be demonstrated further in Chapter 3 and previous studies (Aldayel and Magdy, 2019a; Sobhani, 2017) to be sub-optimal. A recent study by (Xi et al., 2020) analyzed the users posted images in Facebook to

understand the ideological leaning and how political ideology is conveyed through images. In their work, they used the scores generated by CNN image classifier model to further explore the features that distinguish Liberals from the Conservatives. The most distinct features predicting liberal images are related to economic equality. While for conservative members tend to use images that have objects related to state and economic power such as “court”.

On the other hand, stance detection has been used to analyze the attitude related to disruptive events (Demszky et al., 2019; Darwish et al., 2018). For instance, the work of (Darwish et al., 2018) studied the public attitude towards Muslims after Paris terrorist attacks in 2015. In their work they collected tweets that mentions keywords related to Islam. After that, they analyzed users’ interactions and used the historical tweets to predict their attitudes towards Muslims. The work of (Demszky et al., 2019) analyzed the attitudes about 21 mass shooting event. They first derived list of mass-shootings events between 2015 and 2018 from the Gun-Violence Archive. For each event, they defined a list of key words to collect tweets related to these events. Then to study the polarization of opinion toward these events they estimated the party for each user based on the political accounts they follow.

2.6.2 Applications related to social-based phenomena

Stance detection has been used to solve algorithmic issues on social media platforms. These issues are a reflection of social phenomena on these platforms. The most common phenomena that affect various social media platforms are echochambers and homophily (Fuchs, 2018). Previous studies have concluded that social media are polarized in nature which boost homophily behavior (Barberá et al., 2015; Bessi et al., 2016; Quattrociocchi, Scala, and Sunstein, 2016; Darwish, Magdy, and Zanouda, 2017a; Garimella and others, 2018). Homophily is the social phenomenon that concerns with people tendency to connect with “like minded friends”. Echo-chamber is the cascade of a certain information among group of people. This social behavior has been magnified in social media structures where certain beliefs are amplified within close circles of communication. Consequently, people are exposed to content with consent to the same opinion that they hold. As a result, this reinforce social media users’ views biases and blinds the users from other sides of information. Therefore, stance detection has been used to help in measuring and alleviating the problems resulted from polarization on social media. For example, the study by (Garimella et al.,

2017) uses stance as the main factor to expose the user to contradicting views. Moreover, stance detection was used to identify and measure the controversy level on the social media platform (Al-Ayyoub et al., 2018; Dori-Hacohen and Allan, 2015; Jang and Allan, 2018). In the study by (Jang and Allan, 2018) they used the stance summarization technique to rerank controversial topics. Their method collects arguments on five topics and summarize the overall stance with respect to top tweets for a given topic.

2.6.3 Veracity checking applications

The rapid dissemination of news on social media platforms encourages people to depend on these platforms as the main source of information. This kind of information consumption triggers critical issues in social media related to the credibility of the exchanged information, namely, fake news and rumour detection. Recently, a surge of attention has been devoted to classifying stances to help in setting the first step towards solving veracity checking issue (Derczynski et al., 2017b; Allcott and Gentzkow, 2017).

A rumour is defined as an uncertain piece of information that lacks a secure standard of evidence for determining whether it is true or false (Levinson and Ember, 1996). It has been shown that false rumours have an obvious effect across various fields. For example, in the healthcare sphere, false rumours on Web pose a major public health concern. In 2014, rumours about the Ebola epidemic in West Africa emerged on social media, which made it more difficult for healthcare workers to combat the outbreak (Shultz, Baingana, and Neria, 2014). Rumours also affect modern-day journalism, which depends on social media as the main platform for breaking news. Additionally, there have been numerous occasions where false rumours have yielded severe stock market consequences (Schmidt, 2015). To this end, many initiatives, such as Emergent dataset, PHEME dataset and RumourEval SemEval-2017 and 2019 datasets (Ferreira and Vlachos, 2016; Derczynski et al., 2017b; Kochkina, Liakata, and Augenstein, 2017; Derczynski et al., 2017a). These initiatives have been conducted to encourage the development of tools that can help with verifying misinformation in news articles.

For these task, stance detection has been used as a key feature for checking the credibility of a piece of news. As discussed in section 2.3.1, stance in comments towards the news is measured to detect if these comments are confirming or denying

the news, which is used later to detect if the news is a rumour or authentic.

Unlike rumours, fake news aims to create a misleading information and it is always false (Zubiaga et al., 2017). Data veracity refers to the truthfulness, trustworthiness, and accuracy of the content. Social media posts have massive amounts of user-generated contents through which fake information can be easily spread. One particularly challenging task is verifying the truthfulness of social media information, as the content by nature tends to be short and limited in context; these characteristics make it difficult to estimate the truthfulness of social media information compared to other information resources, such as news articles. One of the proposed methods to address fake news is based on detecting the stances towards news organizations. This is due to the fact that understanding other organizations stances toward a news topic helps in inferring the truthfulness of a news article. Fake News Challenge initiative (FNC-1) adopted this approach and proposed a stance detection task to estimate the stance of articles to a given headline (claim). The best performing system in the Fake News Challenge was proposed by (?). In their study, they used gradient-boosted decision trees and convolutional neural network (CNN) along with many textual features. Another recent study by (Mohtarami et al., 2018) achieved relatively similar results to the best system with feature-light memory network model enhanced with (LSTM and CNN). A more recent study by (Shu, Wang, and Liu, 2019) used three features extracted from user’s interactions, news author and contents of a news article to better detect the fake-news. In the study by (Ghanem, Rosso, and Rangel, 2018), they used a combination of lexical, word embedding and n-gram to detect stance in two datasets (FNC-1) and Emergent. A recent study by (Borges, Martins, and Calado, 2019), proposed a new text representation that incorporates the first two sentences of the article along with the news headline and the entire article to train a bidirectional RNN model using (FNC-1) dataset.

2.7 Stance Detection Resources

This section lists the current available resources for stance detection tasks. Tables A.3 and A.4 in Appendix A present in chronological order the available datasets that have been annotated with stance labels. These tables categorise the datasets on a higher level as classification and prediction datasets, as defined in section 2.3.2. In the classification tasks the datasets are further categorized as: target-specific, multi-target and claim-based stance dataset. For the stance prediction datasets, they are further catego-

rized as macro and micro predictions.

Target-specific datasets: There are five publicly available datasets that contain stance annotations for predefined targets on social media, table A.3 in Appendix A. The first dataset is the SemEval stance detection Mohammad et al. (2016b), which provides two sub-datasets to serve two frameworks: supervised framework (Task A) and a weakly supervised framework (Task B), that we have discussed both earlier in section 2.5.1. Another work by Gautam et al. (2019) provides a dataset related to (*Me Too*) movement. This dataset contains around 9000 tweets annotated with stance, hate-speech, relevance, stance, dialogue, act and sarcasm.

Most of the target specific stance detection datasets in social media are English sources. There are two distinct stance datasets that covers non-English stance in social media. The first dataset is the MultistanceCat dataset Taulé et al. (2018b), which contains tweets related to Catalan Referendum in Spanish and Catalan. The dataset provides a multi-modeling to the stance in social media by incorporating the information included in the link along with the text of the tweet. The other dataset is the "*SardiStance*" which is related to *Sardines movement* in Italian tweets. This dataset has been introduced as part of EVALITA2020 task Cignarella et al. (2020). This task provides two variations of data based on two subtasks (a) Textual Stance Detection and (b) Contextual Stance Detection. For the Contextual Stance Detection, the dataset contains a wide range of contextual information related to the post level, such as number of retweets and number of replies, along with data related to social network of user level such as friends, replies and quotes' relations.

Claim-based datasets: In this kind of stance dataset the object of evaluation is the source of information instead of a social actor. The Rumours dataset Qazvinian et al. (2011) is a claim-based stance detection dataset designed for Rumours resolutions. This dataset contains 10,417 tweets related to Obama, Air, France, cellphone, Michelle, and plain. In this dataset the rumour tweet is evaluated against set of other tweets to define the stance of these tweets in supporting or denying the source of the rumour. Additionally, the Emergent dataset Ferreira and Vlachos (2016) is a Claim-based stance detection dataset for Fact checking. This dataset contains rumours from a variety of sources such as rumour sites, e.g., snopes.com, and Twitter. Another dataset available for claim detection is Fake-News dataset. This dataset contains news articles from the Emergent dataset where the news headline is being evaluated against set of body text.

The SemEval 2019 rumours detection dataset by (Derczynski et al., 2017a), en-

riched the SemEval 2017 (rumours detection task) dataset by adding new data from Reddit and extend the language representation of this dataset to include Russia as new topic. Moreover, a recent study by Conforti et al. (2020) provided a dataset called "Will-They-Won't-They" (WT-WT), for a rumor verification task that contains around 51K tweets covering the financial domain.

Multi-related-targets: The two datasets that have multi-related-targets stance annotations are the Trump vs. Hillary dataset Darwish, Magdy, and Zanoouda (2017b) and Multi-targets dataset Sobhani, Inkpen, and Zhu (2017). In Trump vs. Hillary dataset, each tweet is a stance annotated for the two candidates in the same time such as (supporting Hillary and Against Trump). The same annotation technique has been used in Multi-targets dataset for an extended list of US presidential candidates. The Multi-target dataset Sobhani, Inkpen, and Zhu (2017, 2019) contains three pairs of targets Clinton-Sander, Clinton-Trump, and Cruz-Trump.

Stance prediction datasets: As a result of the lack of benchmarks datasets in this kind of stance detection, the researchers tend to build their own datasets as illustrated in A.2, table A.4. These datasets are constructed to predict the stance before the event time. For instance the dataset by Darwish, Magdy, and Zanoouda (2017a), used twitter as a source for stance prediction data, which is the only available dataset for this type of stance detection in social media. The other two datasets by Qiu et al. (2015) Dong et al. (2017) use data collected from online forms to develop a stance prediction model to infer the users stance based on historical contextual data.

2.8 Current trends

It is worth noticing that a small amount of work used stance to analyze social issues in comparison with political topics. This is due to the controversial nature of the political topics which facilitates the data collection for stance detection.

Stance detection has been mostly approached using classification-based algorithm. This is mostly applied by using supervised learning algorithms with large dependency on human-annotated data. Consequently, techniques such as transfer learning and unsupervised learning have been used to resolve the scarcity of annotated data but with less attention from researchers compared to supervised methods. This scarcity reflected by the need to enrich the data with information related to the object of interest. For instance, to detect stances related to climate change, information related to global warming considered beneficial for stance detection in-order to cover the complete as-

pect of the topic. Other studies worked on overcoming this issue by using distant supervision to annotate the data without having to manually label the data.

On the other end of the spectrum, few studies have dealt with stance prediction, which is mostly handled in relation to specific events. The main goal behind this kind of stance detection is to predict the unexpressed views and to infer people standpoints on an event in advance (pre-event). Therefore, stance prediction suits the analytical studies to examine the temporal effects of various events on public opinion such as candidates, elections (Darwish, Magdy, and Zanouda, 2017b) or social phenomena as Islamophobic (Darwish et al., 2018). In this kind of studies, the dataset contains pre-event and post-event posts annotated with the user's stance before and after the event consequently. Thereby, predicting the stances is based on the user's past behavior which can be extracted from network features along with the post's content (Himelboim, McCreery, and Smith, 2013).

While there has been a large interest from the NLP community on developing effective approaches for stance detection Küçük and Can (2020) that mainly modeled the task as a text-entailment task, there is also a large amount of work from the social computing and computational social science communities that showed the effectiveness of using user's interactions and network feature for stance detection and prediction. This shows that the task of stance detection is a multidisciplinary task that has been of interest to multiple computer science communities.

2.9 Summary

This chapter showed an overview of stance modeling on social media. First, we explained the different types of targets when applying stance detection, namely: target-specific, multi-related targets, and claim-based. Then we showed the existing work on detecting the expressed stance or predicting unexpressed user's stance on future events. Later, the most used features for modeling stance and different machine learning approaches were discussed and compared, showing that network features are superior to content (textual features) for most of the studies on stance detection. Moreover, supervised methods using SVM were found to be the most effective for different datasets. In addition, the recent attempts for applying transfer learning and unsupervised learning for stance detection have promising results and is expected to be one of the main research direction in this area in the future. We also discussed the different applications of stance detection, including miss-information detection which is one of the most pop-

ular research topics in the recent couple of years. Finally, we summarise the existing resources on stance detection, including datasets and most effective approaches.

The next chapter examines the association between sentiment and stance, and it evaluates the effectiveness of sentiment analysis methods for stance detection.

Chapter 3

Evaluation of sentiment as stance

In Chapter 2, we showed the tendency of several studies to model the stance on social media as a textual entailment task, where for a given text, the aim is to infer the in-favor and against stances toward a given topic. This kind of modeling has been introduced by the work of (Mohammad et al., 2016b), which provides a dataset that contains sentiment along with stance labels to experiment with possible ways to leverage the sentiment polarity to enhance the stance detection methods. Since then, there has been a noticeable misconception in social media analysis studies between sentiment and stance. Some of the previous studies tend to use sentiment analyzers as the main method to measure the support of a given target. This is due to the ubiquity of sentiment analyzers, which attract most of the researchers to utilize sentiment as the only means to gauge the support towards a given topic. Thus, in this chapter, we answer the first research question of the thesis *”RQ1: Can sentiment polarity be used to identify the stance towards an event? What is the association between sentiment polarity and the stance?”* Particularly, we highlight the main theoretical differences between the stance and sentiment and illustrate these differences with some examples from two stance datasets. Moreover, we evaluate the practice of leveraging the sentiment features as a predictor of the stance by examining the efficiency of eight different sentiment analysis models in detecting the stance and comparing their results with the stance detection model.

3.1 Introduction

(Ochs, 1996) defines the affective stance as *”a mood, attitude, feeling and degree of emotional intensity”*. Many studies have explored the domain of emotion polarity

(sentiment) as they arise in interaction as a proxy for stance and attitudes. Sentiment analysis is a well known task in NLP to determine the polarity of emotion in a piece of text. Generally, this task could be defined as estimating individual emotional polarity, as either being positive, negative, or neutral (Pang and Lee, 2008; Jurafsky and Martin, 2008). This can be seen in the work of SemEval sentiment tasks SemEval-(2015, 2016, 2017 and 2020), where the aim of the sentiment analysis task is to determine the polarity towards an aspect of a product such as cell phone (Patwa et al., 2020; Pontiki et al., 2014a; Nakov et al., 2013a; Pontiki et al., 2014b; Nakov et al., 2013b).

The stance can be defined as the expression of the individual's standpoint toward a proposition (Biber and Finegan, 1988). Detecting the stance towards an event is a sophisticated process where various factors play a role in discovering the viewpoint, including personal and social aspects. Most of the studies in this area have focused on using the textual elements of the social media user's posts such as sentiment of the text to infer the stance (Somasundaran and Wiebe, 2010; Elfardy and Diab, 2016; Ebrahimi, Dou, and Lowd, 2016). While the goal of the stance detection is to determine the favorability towards a given entity or topic (Mohammad, Sobhani, and Kiritchenko, 2017).

There is a large body of research where the sentiment has been used solely to discover the viewpoints towards an event (Lee, 2018; Overbey et al., 2017; Unankard et al., 2014; Tsolmon, Kwon, and Lee, 2012). These studies hypothesised that the sentiment polarity could indicate the stance. However, another line of research develops a stance specific model to infer the viewpoints where sentiment is being neglected (Darwish et al., 2020b; Darwish, Magdy, and Zanoouda, 2017a; Trabelsi and Zaïane, 2018). As the dependence on sentiment as a sole factor for stance, the prediction has been found to be suboptimal, which might indicate a weak relation between sentiment and stance (Mohammad, Sobhani, and Kiritchenko, 2017; Elfardy and Diab, 2016).

Accordingly, it becomes important to examine the relation between sentiment and the stance for identifying the viewpoint towards an event.

Research questions. To answer our first main research question in this chapter (RQ1 of the thesis), we propose the following subresearch questions:

- RQ1.1: Can sentiment polarity be used to identify the stance towards an event?
- RQ1.2: How does sentiment align with stance? When does positive/negative sentiment indicate support/against stance?
- RQ1.3: How effective sentiment analysis methods are in detecting stance?

These questions aim to identify whether the sentiment can substitute the stance by studying the polarity of the expressed stance on a fine-grind tweet level. In other words, this chapter examines whether the supporting/opposing stances can be identified with positive/negative sentiment. Moreover, we evaluate the effectiveness of the sentiment analysis method in detecting the stance. To answer these questions, we used the SemEval stance dataset (Mohammad et al., 2016b), the popular stance dataset that contains sentiment and stance labels. To further validate the results, we constructed a new stance detection dataset with about 6000 tweets towards four topics and annotated it with gold labels for sentiment and stance. This dataset contains the parent tweets along with replies tweets, which provides contextualized information for the annotator and helps in judging the sentiment and stance of the replies tweets. After that, we analyze the datasets to determine the correlation between sentiment polarity and the gold label stance.

In the following sections, an analysis of the association between the stance and the sentiment towards set of targets/topics is applied to measure the alignment between sentiment polarity and the expressed stance. Furthermore, we provide a comparative analysis between eight sentiment analysis models and the stance detection model to highlight the efficiency of sentiment analysis as a stance indicator.

3.2 Tasks definitions

In general, sentiment analysis concerns with detecting the polarity of the text, which can be inferred without the need of having a given target of interest; for example "*I am happy*". Thus, sentiment analysis model can be represented as shown in equation 3.1, where T is a piece of text, and the outcome is usually one of three labels $\{positive, negative, neutral\}$. However, the main sentiment outcome can take different forms such as binary polarity, multi-level polarity, and regression.

$$Sentiment(T) = \{Positive, Negative, Neutral\} \quad (3.1)$$

Another kind of sentiment analysis work concerns with inferring the polarity of a text by using a predefined set of targets, this kind of sentiment analysis is usually referred to as target-based sentiment analysis (Ma, Peng, and Cambria, 2018; Pontiki et al., 2014b; Karamibekr and Ghorbani, 2012; Singh, Singh, and Paul, 2015). In this

kind of sentiment analyses, the classification task can be defined as follows:

$$\textit{Sentiment}(T|G) = \{\textit{Positive}, \textit{Negative}, \textit{Neutral}\} \quad (3.2)$$

In definition 3.2, the input G represents the target or the entity of evaluation. Still, in this kind of sentiment analysis, the dominant factor to gauge the polarity is the raw text. As demonstrated in Chapter 2, section 2.3.1 a separate stance classification model needs to be built for each target (G), unlike sentiment where a general model can be trained for a target-independent sentiment analysis.

3.3 Association between sentiment and stance in literature

In the literature, sentiment has been widely used either to infer public opinion or as a factor to help in detecting the stance towards an event. This section discusses these cases with a focus on studying the stance towards an event where the simple sentiment has been used either by using a sentiment lexicon or the textual polarity of the text.

3.3.1 Sentiment as stance

Sentiment has been used interchangeably with stance to indicate the viewpoint detection (Park et al., 2011; Hu, Wang, and Kambhampati, 2013; Smith et al., 2017; Lee, 2018; Unankard et al., 2014; Tsolmon, Kwon, and Lee, 2012; Agarwal, Singh, and Toshniwal, 2018). In these studies, the sentiment polarity has been used purely as the only factor to detect the viewpoint towards various events in social media. For instance, the work of (Smith et al., 2017) used sentiment to investigate the opinion towards the terrorist attack in Paris, during November 2015. They used annotators from (Appen) to label the sentiment (negative, positive, or neutral) as expressed in the tweet and used these labels as a way to analyse the public reaction toward Paris attacks in 2015. In a study done by (Park et al., 2011), they used the sentiment to discover the political leaning of the user's comments on news articles. In their study, a sentiment profile was constructed for each commenter to help in tracking their polarity toward a political party. For instance, a liberal commenter uses negative comments in conservative articles and positive comments to liberal articles.

A more recent study by (Lee, 2018) used the sentiment to examine the opinions following the release of James Comey's letter to Congress before the 2016 US

presidential election day. The previous study categorized 25 most common hashtags with sentiment polarity towards Hillary Clinton and Trump. Furthermore, the work of (Unankard et al., 2014) used sentiment analysis to analyze the political preferences of the users for the 2013 Australian federal election event. For the sentiment, they recruited three annotators to label the tweet with a polarity score (positive, negative, or neutral). In their study, they used aspect-level sentiment for predicting the user's political preference and they overlooked the cases where the sentiment is negative and the stance is expressing a support viewpoint.

Another study (Tsolmon, Kwon, and Lee, 2012) developed an opinion equation based on sentiment lexicon and frequency of a term to infer the users' opinions towards events as they extracted from the timeline. In addition, the work of (Hu, Wang, and Kambhampati, 2013) designed topic-sentiment matrix to infer the crowd's opinion. Another recent study by (Agarwal, Singh, and Toshniwal, 2018) used AFINN-111 dictionary for sentiment analysis and used sentiment polarity as an indication of opinion towards Brexit. All the above studies treated sentiment as the indicator of the stance toward the event of the analysis.

3.3.2 Sentiment as a proxy for stance

Another line of research used sentiment as a feature to predict stance (Somasundaran and Wiebe, 2010; Elfardy and Diab, 2016; Ebrahimi, Dou, and Lowd, 2016; Mohammad, Sobhani, and Kiritchenko, 2017). In the popular SemEval stance dataset (Mohammad et al., 2016b), the tweets are labeled with sentiment and stance to provide a public benchmark to evaluate the stance detection system. In their work, they showed that sentiment features are useful for stance classification when they are combined with other features and not used alone. The work of (Ebrahimi, Dou, and Lowd, 2016) used an undirected graphical model that leverages interactions between sentiment and the target of stance to predict the stance. Also, the work of (Somasundaran and Wiebe, 2010) developed a stance classifier that used sentiment and arguing expressions by using the sentiment lexicon along with the argument lexicon which outperforms Unigram feature system. In (Igarashi et al., 2016) they used SentiWordNet to produce sentiment for each word and use the sentiment value along with other features to predict the stance in SemEval stance dataset and compared with CNN stance model. They found that feature based model performed better in detecting stance. The work of (Krejzl and Steinberger, 2016) used surface-level, sentiment and domain-specific fea-

tures to predict stance on SemEval 2016 stance dataset. Overall, the use of sentiment in conjunction with other features helps in predicting the stance but not as the only dependent feature.

The work of (Mohammad, Sobhani, and Kiritchenko, 2017; Sobhani, Mohammad, and Kiritchenko, 2016) examined the extent to which sentiment is correlated with stance in the sense of enhancing stance classifiers. The main focus of the previous study was to investigate the best features for the stance classification model. In their work, they concluded that sentiment might be beneficial for stance classification, but when it is combined with other factors.

This chapter provides a thorough examination of the sentiment-stance association with focus on gauging the alignment between sentiment and stance by analysing in depth the relation of how the stance is being expressed in conjunction with the sentiment.

3.4 Stance and sentiment datasets

In this section, we describe the stance datasets that has been used to investigate the sentiment polarity of the expressed stance.

3.4.1 SemEval stance dataset

We study the sentiment polarity in the expressed stance. To accomplish this, we used SemEval stance dataset which contains about 4000 tweets on five topics, including Atheism (A), Climate Change (CC), the Feminist Movement (FM), Hillary Clinton (HC) and the Legalisation of Abortion (LA). This dataset contains sentiment polarity labels along with stance labels. The distribution of stance and sentiment is shown in figure Figure 3.1. This dataset contains a tweet–target pairs annotated for both stance and sentiment. The tweet may includes an explicit or implicit indication of the target of viewpoint in the tweet (the target is not directly mentioned). This can be seen as about 26.5% of the tweets contain a target of the viewpoint, that is someone/something other than the target itself. About 28% of the ‘Hillary Clinton’ tweets and about 67% of the ‘Legalization of Abortion’ tweets were found to have an implicit target, where the tweets do not mention ‘Hillary’ or ‘Clinton’ and in the case of (LA) they do not mention ‘abortion’, ‘pro-life’, and ‘pro-choice’. This dataset has been introduced in SemEval-2016 task competition with 19 teams submissions. The task concluded with

providing a simple baseline system that outperforms all 19 teams that participated in the shared task (Mohammad et al., 2016b).

3.4.2 Context-dependent dataset

We constructed a context-dependent (CD) stance dataset that contains 6324 reply tweets covering four controversial topics: Antisemitic (AS), Gender (G), Immigration (I), LGBTQ (L). Table 3.1 shows the distribution of the tweets with respect to each topic. In this dataset, each tweet has been annotated by five annotators using Appen¹, and the label with a majority vote is assigned. We used the same annotation guideline of SemEval stance dataset (Mohammad et al., 2016b). Since CD dataset is all reply tweets, the parent tweets along with reply tweets have been provided to the annotators to understand the context of the conversation to better judge the sentiment and stance. This dataset contains tweet–target pairs, where each tweet has two labels indicating the stance and the sentiment polarity of the tweet. Figure 3.1 illustrates the overall distribution of stance labels concerning the sentiment and stance for the four topics.

SemEval stance	#	CD stance	#
Atheism (A)	733	Antisemitic (AS)	1050
Climate Change is Concern (CC)	564	Gender (G)	1050
Feminist Movement (FM)	949	Immigration (I)	3174
Hillary Clinton (HC)	934	LGBTQ (L)	1050
Legalization of Abortion (LA)	883		
Total	4063	Total	6324

Table 3.1: Number of tweets for each topic.

3.5 Correlation between sentiment and stance

3.5.1 Agreement between sentiment and stance labels

To gain a good insight of how the stance is being expressed, we first analyze the distribution of stances and sentiment on the topic level. Figures 3.1 a and b, illustrate the stance and sentiment distribution in the SemEval stance dataset and CD stance dataset, respectively. Overall, the negative sentiment constitutes the major polarity of the most

¹<https://appen.com/solutions/annotation-capabilities/>

topics. This reveals the tendency of using negative sentiments to express a viewpoint on a controversial topic. It can be observed that for climate change the supporting stance constitutes about 59%; however the overall tweets with negative sentiment constitute 50%. Furthermore, 30% of the LGBTQ tweets show negative sentiment, while only 7% of the tweets express the opposing stance. This results demonstrate that sentiment does not simply represent the stance in social media posts.

Figure 3.2 illustrates the sentiment distribution over the stance in the two datasets. The graphs show that the negative sentiment constitutes the major polarity over the Favor and Against stances. As the negative sentiment represents over 56% and 54% of the supporting stance in the SemEval and CD stance datasets, respectively. These results reveal the tendency of using negative sentiments to express a viewpoint towards a controversial topic. Moreover, we analyzed whether sentiment and stance are independent of each other. We used Cramer's test (Cramér, 1999) to gauge the strength of relationship between sentiment and stance. The result from Cramer's test indicate that the variance (V) value ranges between 0 and 1, with 1 as an indication of a high association between the nominal variables (Liebetrau, 1983). By applying Cramer's test in the SemEval stance dataset, the resultant " V " value = 0.12, which is a strong indication of the independence of sentiment and stance from each other. The same test applied to CD dataset and show the independency between stance and sentiment with " V " value = 0.10.

Table 3.2 shows some examples where the sentiment does not associate with the stance. Examples 1 and 2 show tweets with an opposing viewpoint on targets, while using positive and neutral sentiment. Examples 3 and 4 show the opposite situation,

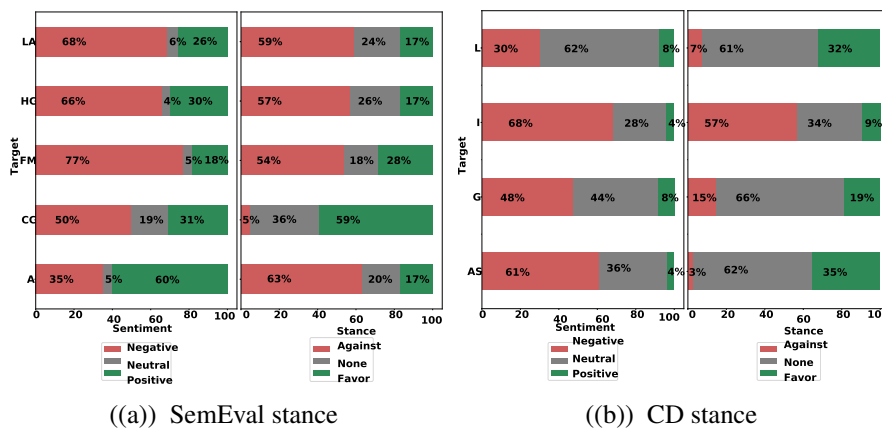


Figure 3.1: The distribution of sentiment and stance with respect to each topic.

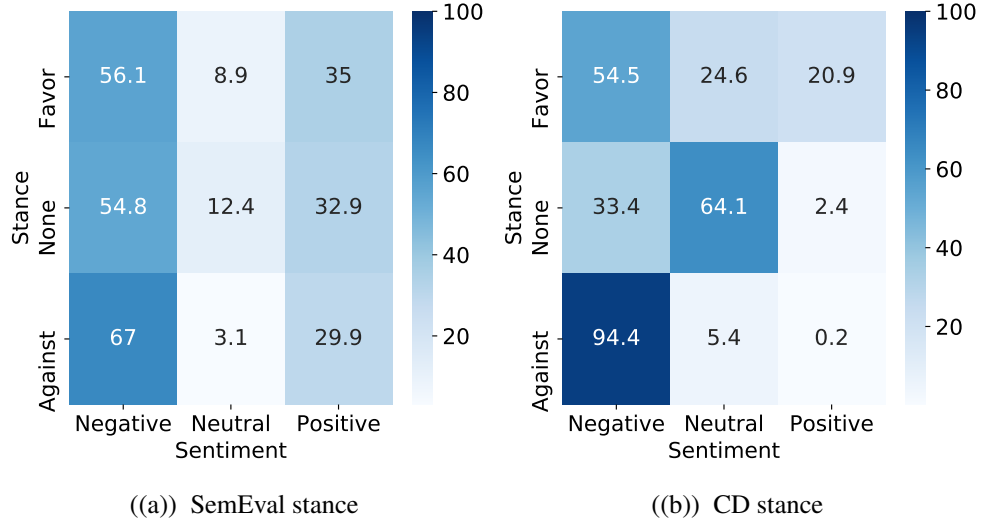


Figure 3.2: Distribution of sentiment per a given stance.

where the expressed stance is supporting, while the sentiment is negative.

3.5.2 Analysis of the textual patterns

To gauge the similarity between the vocabulary choice that has been used to express sentiment and stance, we analyzed the tweets in the two datasets using Jaccard similarity. We used Jaccard coefficient, the widely adopted measure to capture the overlap between two sets An et al. (2019); Achananuparp, Hu, and Shen (2008); Gomaa and Fahmy (2013). In this analysis, for each sentiment and stance labels, we combine all tweets and use Term Frequency-Inverse Document (TF-IDF), to find important words in each type of sentiment and stance. In order to compute the TF-IDF on the tweet level we consider each tweet as document. Using TF-IDF helps in filtering out less significant words. The Jaccard similarity between the set of sentiment and stance words is defined as following:

$$Jaccard(W_{sentiment}, W_{stance}) = \frac{W_{sentiment} \cap W_{stance}}{W_{sentiment} \cup W_{stance}} \quad (3.3)$$

Where $W_{sentiment}$ and W_{stance} denote the list of top N words by TF-IDF value for the tweets with specific sentiment and stance type.

Fig 3.3 shows that the similarity between the words that have been used to express favor stance has less than 20% of similarity with tweets that has a positive sentiment. That means users tend to express their Favor stance without using positive sentiment words. In contrast, the common words for against stance have the most significant

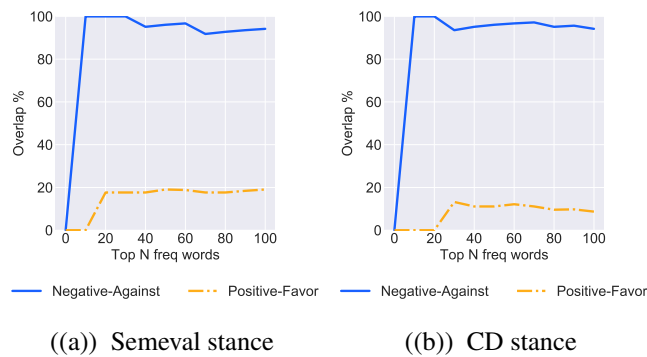


Figure 3.3: Jaccard similarity of the top N-most frequent words between sentiment and stance.

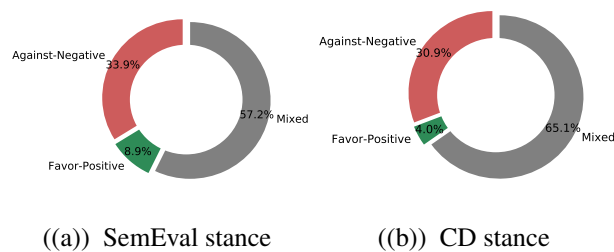


Figure 3.4: Tweets with matching and mixed stance and sentiment.

similarity with against sentiment words. The Jaccard similarity becomes stable with growing N. As Fig 3.4 shows that the overall agreement between the sentiment and the stance is minuscule in general. The tweets that have against-negative labels constitute less than 33%. Similarly, less than 8% of the data has positive sentiment and favor stance. This shows that in general, negative words tend to be similar to the against words, while the matching cases are minuscule. On the other hand, the matching cases where the tweet expresses favor and positive sentiment constitute about 8.9% and 4% of the overall data of SemEval stance and CD stance datasets.

These results show that sentiment is sub-optimal in inferring the real stance on a topic. There is a clear mismatching between the negative/positive sentiment and the supporting/against stance. Even with the dominance of the negative sentiment in most of the topics, yet the overall stance has shown a mixer of support viewpoint.

3.5.3 The association between sentiment and stance

In order to answer the first research question "RQ1.1", the previous analysis of the association between sentiment and stance shows that the sentiment cannot substitute

the stance in general. The word choice shows a gap exists for in-favor stance and positive sentiment. Hence, using sentiment polarity as the only factor to predict public opinion potentially leads to misleading results. The result of the mismatch between in-favor and positive stance was sizable. As the positive sentiment does not usually represent supporter viewpoints.

As for the overall alignment between sentiment and stances "RQ1.2", there is a noticeable dis-alignment between sentiment and stance for a given topic. In general, the sentiment tends to be negative in the expressed stance as a way to rebuttal or defend the viewpoint and show support or opposing stance. The negative sentiment could help in discovering some of the against stances, but it will be mixed with a proportion of the supporter viewpoints.

3.6 Sentiment as predictor of the stance

In this section, we verify the effectiveness of sentiment analysis for stance prediction by evaluating the effectiveness of sentiment analysis methods in comparison with stance detection. Particularly, we show how well sentiment analysis methods in capturing the support and against stance across two datasets spanning nine topics. We experiment with custom in-domain sentiment analysis (*CS*), along with the general out-domain trained sentiment analysis methods (*GN*). We extended the effort of the study by (Sen, Flöck, and Wagner, 2020). In their study, they analyzed the efficiency of sentiment analysis in predicting the stance towards political figures. In this study we extended the analysis to verify the effectiveness of sentiment on predicting the stance of various topics converging different domains.

3.6.1 Sentiment models

We use two types of sentiment analysis methods: out-domain (*GN*) and in-domain (*CS*) target specific sentiment analysis. For in-domain sentiment analysis, we use Aspect based sentiment model TD-LSTM (Tang, Qin, and Liu, 2016). They used syntactic dependencies and trained LSTM on two datasets from SemEval 2014 and achieved state-of-the-art performance. For the out-domain models, we implemented four well-known sentiment analysis. We use VADER (Hutto and Gilbert, 2014), use lexicon along with rule based model to understand the syntactic characteristics of a sentence such as negation. Moreover, we used MPQA (Hu and Liu, 2004), which pro-

Table 3.2: Sample of tweets illustrating the sentiment (Sent) polarity of the expressed stance (Stan). The examples are collected from SemEval stance and CD datasets.

#	Tweet	Target	Sent	Stan
1	It is so much fun having younger friends who are expecting babies. #beentheredonethat #chooselife .	Legalisation of Abortion	+	-
2	Life is sacred on all levels. Abortion does not compute with my philosophy. (Red on #OITNB)	Legalization of Abortion	0	-
3	The biggest terror threat in the World is climate change #drought #floods	Climate Change is the real concern	-	+
4	Apparently, @BernieSanders fans don't like to see Hillary's name even implied. #SheWhoShallNotBeNamedBecauseShesWinning	Hillary Clinton	-	+
5	Thank you Sen. Graham for your support of @POTUS and his rational immigration and border-enforcement policies!	Immigrants	+	-
6	@realDonaldTrump No wall - don't need it ; not paying for it Your hype is not working! There are better ways to work on this situation than a wall and shooting people at the border #CorruptGOP	Immigrants	-	+
7	That's why I blocked him. That one comment! Tired of men thinking women can't do the job!	Gender	-	+
8	Wow, it really scares you that much that women are finally getting a say?	Gender	0	+
9	We should be civil in our discourse when fighting back against bigotry, xenophobia and anti semitism. Jews cannot allow history to repeat itself in America.	Anti- semitism	0	+
10	There you go, dividing again. Thanks for letting real abusers off the hook too. Go after them and not the trans community.	LGBTQ	-	+
11	If I actually believed that you wanted a Brexit deal I might agree but the reality is you want to sabotage Brexit and keep us in the EU	Brexit	-	+

posed feature-based opinion summarization which depends on a feature terms' occurrence counts heuristics. LabMT (Dodds et al., 2011) proposed post-specific sentiment analysis method by constructing a lexicon of 10,000 words. Similarly, SentiStrength (Thelwall, 2017) used lexical approach and rule-based methods to deal with linguistic features of a given post.

3.6.2 Evaluation metric

For evaluating the effectiveness of sentiment on predicting the stance, we use the SemEval stance detection official evaluation script to calculate the F1-score (Mohammad

et al., 2016b). The macro F_{avg} is calculated as presented in equation 1.

$$F_{avg} = \frac{F_{favor} + F_{against}}{2} \quad (3.4)$$

(1)where F_{favor} and $F_{against}$ are calculated as shown below:

$$F_{favor} = \frac{2P_{favor}R_{favor}}{P_{favor} + R_{favor}} \quad (3.5)$$

$$F_{against} = \frac{2P_{against}R_{against}}{P_{against} + R_{against}} \quad (3.6)$$

Note that the evaluation measure does not disregard the 'none' class. By taking the average F-score for only the 'favor' and 'against' classes, we treat 'none' as a class that is not of interest or 'negative' class in Information Retrieval (IR) terms. Falsely labeling negative class instances still adversely affects the scores of this metric.

3.6.3 Performance of sentiment analysis with stance

To answer the third research question of this study "RQ1.3", the previous experiments show that using sentiment analysis models to predict the stance is suboptimal. We compared eight sentiment models using in-domain (CS) and out-domain (GN) trained models. The findings demonstrator that the general purpose (GN) provides the worst results in comparison with the in-domain trained sentiment analyzer. However, the performance of both types of sentiment analysis models are lower in comparison with the baseline stance detection model. This finding emphasises on the need to enhance the current stance detection models for text-based analysis tasks. The ubiquitous of sentiment analysis models magnified the misuse of these models to predict the stance towards an event or topic. This can be noticed in the performance of sentiment models in predicting the support stance, where the Macro-F1 of Favor class is lower than the Macro-F1 of Against class in the nine topics using the eight models.

In summary, our analysis in this chapter shows how sophisticated is stance detection and that it cannot be simply modeled using the sentiment polarity. This can be seen as the sentiment analyzers are not able to predict the stance in the two datasets. This finding is crucial, especially when assessing the credibility of the results in studies that used sentiment to measure public support of a given topic on social media.

Model	Topic					Overall			Acc
	A	CC	HC	FM	LA	F_{favour}	$F_{against}$	F_{avg}	
Baseline_stance	60.07	42.08	55.54	58.87	60.81	61.17	73.80	67.49	66.53
Senti (Truth)	19.88	27.75	57.07	49.28	42.64	24.52	64.91	44.71	48.83
VADER (GN)	24.66	31.00	44.91	33.91	29.63	26.94	45.15	36.04	35.76
MPQA (GN)	20.84	34.69	36.67	30.46	33.60	31.23	34.90	33.07	30.57
labmt (GN)	22.14	38.27	39.52	30.38	31.72	31.64	39.63	35.63	32.33
SentiS (GN)	25.94	44.26	42.10	46.63	34.92	38.50	46.73	42.62	38.85
TD-LSTM (CS)	34.90	25.53	42.83	37.67	42.22	20.32	66.55	43.43	49.95
MNB (CS)	18.47	20.76	53.59	39.10	39.57	17.88	65.28	41.58	49.07
LR (CS)	19.74	29.76	57.65	50.35	44.62	25.32	65.68	45.50	49.31
BERT (CS)	36.60	00.00	36.64	22.45	41.71	19.76	66.11	42.94	23.50

Table 3.3: Sentiment and stance prediction models on SemEval stance dataset. "GN" indicates general out-domain trained model and "CS" is for custom in-domain trained models.

Model	Topic				Overall			Acc
	AS	G	I	L	F_{favour}	$F_{against}$	F_{avg}	
Baseline_stance	30.37	26.51	54.14	26.04	46.87	71.71	59.29	68.01
Senti(Truth)	45.74	38.60	44.59	36.23	17.86	72.70	45.28	65.96
VADER (GN)	14.47	28.46	36.30	28.15	26.67	40.52	33.59	35.19
MPQA (GN)	22.05	27.15	30.48	31.92	30.13	35.07	32.60	35.13
labmt (GN)	10.45	24.35	32.16	24.47	22.82	37.42	30.12	35.19
SenTiS (GN)	14.91	29.57	33.81	24.61	23.37	40.55	31.96	24.61
TD-LSTM (CS)	02.68	14.61	32.15	08.16	02.03	45.83	23.93	39.13
MNB (CS)	03.05	13.97	37.91	09.86	01.09	56.21	28.65	46.02
LR (CS)	03.17	18.83	40.69	13.05	05.58	56.73	31.16	53.12
BERT (CS)	02.77	15.73	04.38	29.53	25.51	03.86	14.69	20.88

Table 3.4: Sentiment and stance prediction models on CD dataset. "GN" indicates general out-domain trained model and "CS" is for custom in-domain trained models.

3.7 Summary

This chapter examines the association between sentiment and stance. To gauge the extent of this relation, we constructed a new stance dataset with sentiment and stance labels. Then we conducted a textual and quantitative analysis of the expressed stance with respect to the sentiment polarity. Moreover, we examined the effectiveness of the sentiment analysis method to predict the stance using eight different models. To answer the first research question of the thesis "*RQ1: Can sentiment polarity be used to identify the stance towards an event? What is the association between sentiment polarity and the stance?*", in this chapter, we showed that sentiment cannot substitute the stance, as the correlation between the two is weak. Overall, the experimenters showed that sentiment analysis methods tend to be suboptimal in detecting the stance. This finding provides an insight for the social media researcher to be more cautious when it comes to identifying the viewpoints on an event and to take into account the clear difference between sentiment and stance. As using sentiment purely overshadows the real stance and leads to truncated results. This finding illustrates the complexity to model the stance on social media which goes beyond the simple sentiment polarity.

The next chapter examines the multi-modality of stance detection as motivated by the sociolinguistic theory by (Bassiouney, 2015). We provide an assessment to different kind of interaction features to model the online stance in social media. As described in Chapter 2, we examine these features and provide a thorough analysis of the online interactions and content features to detect the online stance in social media.

Chapter 4

Possible factors for stance detection on social media

The previous two chapters provides an extensive evaluation of stance modeling on social media. Chapter 3 illustrates the weak relationship between sentiment and stance. As Chapter 2 illustrates that most of the stance detection studies have focused on using the textual elements of the user's posts independently from user behavioural data, such as homophily and network structure. This chapter provides a thorough analysis of stance detection on social media and examines various online factors by evaluating different kinds of interactions. Particularly, this chapter addresses the second research question "*RQ2: To what extent do online interactions predict an individual's stance?*". Examining the effectiveness of social interactions in detecting the online stance provides a better understanding of one of the pressing challenges in predicting viewpoints using only the raw texts of a user's posts. To address this challenge, we assess the modeling of the stance detection on social media by incorporating the online interaction elements and evaluate its effectiveness on predicting the stances towards a topic.

4.1 Introduction

Most of the research on stance detection have modeled the stance as a text classification task, where text of on-topic posts are used as the features (Mohammad et al., 2016b; Elfardy and Diab, 2016; Siddiqua, Chy, and Aono, 2018). Some other work showed the effectiveness of using user's network as the features (Darwish, Magdy, and Zanoouda, 2017a; Magdy et al., 2016; Lai et al., 2016, 2018). However, most of these studies were focused on one topic with no real examination to its generalizability on other topics or

domains. Another limitation of the existing approaches for stance detection is the reliance on signals from active users only who frequently post on social media, where user's stance is modelled either by user's posts or interaction with other users (retweet in case of Twitter). There has been a growing interest on characterizing "silent user" in social media platforms (Bernstein et al., 2013; Gong, Lim, and Zhu, 2015). This group of users known as "lurkers" or "invisible participants" tends to contribute with a little or no content. Some users prefer to interact quietly on social media using other means of interactions instead of directly posting or sharing contents, such as following others and liking posts (Gong, Lim, and Zhu, 2015). Most of stance detection studies used the network representation of the active users only and overlooked the silent users (Darwish, Magdy, and Zanoouda, 2017a; Magdy et al., 2016; Lai et al., 2018).

This study is motivated by du2007stance (Du Bois, 2007) arguments that stance taking is a subjective and inter-subjective phenomenon in which stance-taking process is affected by personal opinion and non-personal factors such as cultural norms. Stance taking is a sophisticated process relates to different personal, cultural and social aspects. For instance, a political stance taking depends on experiential behavior as stated by (McKendrick and Webb, 2014). Thus, users in social media might express their opinion directly by posting about the topic or their stance could be inferred indirectly through their interactions and preferences. Our hypothesis is that user's embedded viewpoint in a post is related to the user's identity which could be better modeled by their interactions and connections in the social network. This idea is related to the concept of homophily in which users with same believes tend to have common interests and group together (Al Zamal, Liu, and Ruths, 2012; Garimella and others, 2018; Darwish, Magdy, and Zanoouda, 2017a).

In this study, we apply an extensive analysis to the possible online signals that can reveal the user's stance. To that end, we examine four groups of signals that might indicate the stance, namely: 1) on-topic posts by the user, which models users who explicitly express their stance on a topic; 2) user's interactions on social media with other users or websites, which models the online social interactions regardless of having the stance expressed or not (IN); 3) user's preferences the posts they like, which enable modeling silent users who do not post or share content only (PN); and finally 4) the network of users they are connected, which enable modeling passive users who might have no content or interaction on social media, but just follow other accounts online (CN). We compare the effectiveness of each of these groups of features on detecting stance individually and when combined. Our main research question is to

understand “What are the factors that can reveal the stance of user online towards a given topic”. We further analyse “how” and “why” these factors might be effective for detecting stance. Particularly, to answer the second main research question of this thesis “RQ2”, we list the following sub-research questions:

- RQ2.1: What are the different signals in a user’s online activity that can reveal their stance, including textual content, networks of interaction (IN), preference (PN), and connection (CN)?
- RQ2.2: Does the performance of detection differ by different types of topics?
- RQ2.3: What makes any of these signals effective (or ineffective) for detecting stance?

Our experiments are applied on the SemEval stance detection benchmark dataset (Mohammad et al., 2016b), which contains a set of over 4,000 tweets labeled by stance towards five different topics. The five topics covers multiple domains not just politics, which makes the dataset ideal to examine the generalisability of the stance detection models, unlike most of work in literature that typically focus on studying one political topic at a time (Lai et al., 2016, 2018; Magdy et al., 2016; Darwish, Magdy, and Zanouda, 2017a). Our results show that training a classification model on pure user network features outperforms the state-of-the-art baseline system (Mohammad et al., 2016b) which is trained on multiple features extracted from the tweets text content. This includes when using the preference network features from only the tweets the user likes and also the connection network of the accounts the user follow, where both can model silent users. When different groups of features are combined, including content and network, a significant improvement is observed. Our findings suggest that for the task of stance detection, even when applied on the level of tweet, user’s network information are more effective features than the content of the tweet itself. This aligns to the sociolinguistic theory in (Bassiouney, 2015), where it defines stance as the link between linguistic forms and social identities which has the capability to establish the alignment between stance-takers. We further applied an extensive analysis to the most influential features for each group of network signals to understand how they outperform textual text. It was interesting to find that the overlap between IN, PN, and CN was not large, where the common nodes among them are around 10% only, however, each of those networks still can model user’s stance towards a given topic. Our analysis to the most influential features from each network on each of the five topics shows

that there is usually some common signals in user online activity that can reveal their stance towards a given topic regardless of the type of the topic. We believe that our findings in this study raises a large concern about protecting the privacy of social media users, where their beliefs and leanings could be easily predicted using any of the footprint signals they leave online. This should motivate social media networks owners and designers to develop methods for protecting the privacy of their users (Waniek et al., 2018).

We have made the collected network information for the SemEval dataset publicly available to allow replication to our experimentation¹.

4.2 Related work

As we discussed in Chapter 2, section 2.4.2 there is a considerable amount of work on viewpoint or stance detection; yet, less work compares the role of content and social actor interactions in stance detection (Himmelboim, McCreery, and Smith, 2013; Darwish, Magdy, and Zanoouda, 2017a). Studying stance on social media needs to cover the intersection dimensions of stance taking process, which are mainly influenced by linguistic forms and social interactions frames (McKendrick and Webb, 2014). Most of the previous studies define stance as a textual entailment task where the main processing depends on the raw text only (Dey, Shrivastava, and Kaushik, 2018; Mohtarami et al., 2018; Augenstein, Vlachos, and Bontcheva, 2016; Mohammad et al., 2016b; Dong et al., 2017). In this form of stance detection, a given text entails a stance towards a premise (target).

It has been shown that constructing a knowledge based dataset about the topic is beneficial in stance detection task (Mohammad et al., 2016b). This constitutes a visible hurdle which limits the stance detection task to a set of predefined topics. Furthermore, many times the topic is not mentioned in the tweet. One way that was suggested to handle the unmentioned target entity in text is to analyze the opinion to the opponent of the entity or supporter of the entity. For example, (Dias and Becker, 2016) constructed a list of keywords that identifies Trump using a dataset labeled with stances toward Hillary. Using this list of keywords help in detecting the unexpressed stand towards Trump. Another study (Dong et al., 2017) follows the same line by constructing corpus that contains words that are *against* and *in-favor* each target to enrich the models. Similarly, (Wei et al., 2016) used a domain corpus related to Trump along with lexicon

¹https://github.com/AbeerAldayel/Stance_detection

to construct a labeled dataset to detect stance towards Trump. Furthermore, (Benton and Dredze, 2018) used context of the users tweets to construct author embedding and predict the stance.

There has been some work on studying the integration of network and content with a limited focus on the ideological political views (Darwish et al., 2020a; Himelboim, McCreery, and Smith, 2013; Lai et al., 2016; Magdy et al., 2016). For instance the study of (Himelboim, McCreery, and Smith, 2013) focused on the liberal and conservative on twitter. Unlike previous work, rather than studying the stance on single topic and using a domain specific data, we study the stance in various domains. This study explores the stance modeling in the social media to know to what extent do network interactions and content interactions reveal an individual's viewpoint. Examining the implications of those interactions in detecting users' stances provides a better understanding of stance modeling on social media. These studies highlights the importance of social network interaction of users to detect their position towards specific events or entities. Nevertheless, they are limited to focusing on one specific topic from the political domain, which lacks examining the generalisability of these approaches on multiple topics from different domains. In addition, it focuses on network interactions that can only model active users who retweet, mention, and reply other accounts.

In this chapter, we use the SemEval benchmark dataset to apply an extensive comparison on stance detection using multiple sets of features and compare it to the state-of-the-art. In addition, we introduce the use of the preference network as a new way to model the stance and examine the possibility of detecting the stance of the silent users. We compare the performance of this new set of features with content-based and other networks based features for stance detection.

4.3 Stance Detection Methodology

In the following, we discuss our proposed methodology including the set of features used and the machine learning method applied. But initially, we discuss the implications of our approach from the conceptual point of view.

4.3.1 User vs Tweet features for Stance Detection

The SemEval dataset is labeled for stance on the tweet level, while in this study we are examining user features in detecting the stance of a given tweet. To enable comparison

to state-of-the-art methods on the same dataset, we apply our detection on the tweet level. This would not be an issue if each tweet in the dataset is coming from a different users. However, we noticed that 167 users (out of 3,528) in the dataset produced multiple tweets. This means that our classifiers trained on network features would always give the same classification to any tweets posted by the same user. We argue that this is acceptable based on the assumption that user stance for a given topic is not expected to change within a short period of time (Borge-Holthoefer et al., 2015).

To further validate our assumption, we examined a set of 167 users who produced multiple tweets on the same topic. Out of those, 104 users have fixed stances in their multiple tweets. The other 42 users have fixed stance on some tweets, and have some tweets with no stance (labeled *none*). Only 19 users have a mix of *favour* and *against* stance in the same topic, but with clear dominance for one of them (e.g. 16 vs 1 tweets). This quick analysis shows that the majority of tweets from the same user are expected to have a fixed stance on a single topic. Thus, we believe that having a fixed set of features, based on user's network, for all tweets of the same user can be seen as an acceptable approach for stance detection on the tweet level.

4.3.2 Feature Extraction

We define four features sets to model the stance in social media. These sets are: on-topic content, user's network interactions, preferences and connections. Those are defined as follow:

- **On-Topic Content (TXT)**, models the text of the tweet, including features combining both word and character n-grams as presented in the best performing system in SemEval 2016 (Mohammad et al., 2016b). This set of features models stance of users who explicitly express it in text.
- **Interaction Network (IN)**, models the network the user interacts with in their posts. It includes the mentioned accounts ($IN_{@}$) and website domains (IN_{DM}) the user interacts with directly either by retweeting, replying, mentioning, or linking.
- **Preference Network (PN)**, models the network the user prefers from the tweets they like. It includes the mentioned accounts ($PN_{@}$) and linked website domains (PN_{DM}) in the tweets the user likes.

Feature Set	Description
TXT	word and character n-grams of the tweet text.
IN:	user’s interaction network. Extracted from user’s <i>Home</i> timeline.
- IN@	the list of accounts the user retweet for, reply to, or mention in their timeline.
- IN _{DM}	the list of web domains the user link in their tweets.
PN:	user’s preference network. Extracted from user’s <i>Likes</i> timeline.
- PN@	the list of accounts mentioned in the tweets the user likes.
- PN _{DM}	the list of web domains in the tweets the user likes.
CN:	user’s connection network. Accounts user connected to.
- CN _{FL}	the list of followers of the user, i.e. accounts that follow the user.
- CN _{FR}	the list of followees/friends, i.e. accounts that the user follows.

Table 4.1: List of feature sets examined in our experiments with their description.

- **Connection Network (CN)**, models the online social ties between the users, which includes the accounts who follow the users (followers CN_{FL}), and those the user follows (friends CN_{FR}).

Table 4.1 shows a detailed explanation of the feature sets. It is worth noting that *IN* features are independent of having users expressing their stance towards the target topic, since it depends on the social and web networks the user interact directly with regardless to the content in tweets. Both *PN* and *CN* features enable modeling silent or passive users who do not post or share content rather than just following or liking tweets from others. Our objective is to understand how each of these feature sets would compare to each other and to the textual features which have been studied heavily in literature.

4.3.3 Stance Detection Model

Since our main contribution is on stance modeling to analyse the effectiveness of different social signals in detecting stance, we used our proposed set of features to train an SVM model with linear kernel for two main reasons: 1) It achieved the best performing model over 19 participating groups at SemEval 2016 (Mohammad et al., 2016b) while outperforming more sophisticated model that used deep learning (Augenstein et al., 2016; Zarrella and Marsh, 2016; Wei et al., 2016). 2) SVM models built with linear kernel are easily to interpret, which would enable us to apply feature analysis for a better understanding to the influential features and their role in stance detection.

Topic	Full dataset		Existing Users	
	Train	Test	Train	Test
Atheism (A)	513 (434)	220 (196)	380 (302)	170 (146)
Climate change is a real concern (CC)	395 (347)	169 (145)	317 (269)	144 (120)
Hillary Clinton (HC)	639 (556)	295 (250)	447 (364)	223 (178)
Feminist movement (FM)	664 (620)	285 (256)	354 (312)	170 (141)
Legalization of abortion (LA)	603 (496)	280 (228)	471 (365)	199 (147)
Total	2814 (2453)	1249 (1075)	1969 (1612)	906 (732)

Table 4.2: Number of tweets used for training and testing with respect to Semeval 2016 topic. The number of unique users authored the tweets are shown in brackets.

We used Scikit-learn² implementation of SVM, which use cross-validation with $k=5$. For all the features, we use vector representation with Boolean value to indicate the presence or absence of the feature’s values. We have examined other feature values, such as frequency and tf-idf, but Boolean values showed the best performance.

4.4 Experimental Setup

4.4.1 Network Features to Detect Unexpressed View

Different sets of user features have been introduced in the previous studies with the focus of defining similar users for specific events as we show in the previous chapter section 2.4.2. These studies highlights the importance of social network interaction of users to detect their position towards specific events or entities. Nevertheless, they are limited to focusing on one specific topic from the political domain, which lacks examining the generalisability of these approaches on multiple topics from different domains. In addition, it focuses on network interactions that can only model active users who retweet, mention, and reply other accounts.

4.4.2 Data Collection

Our experimentation has been applied to the benchmark dataset of the SemEval 2016 stance detection task (Mohammad et al., 2016b). The dataset contains a set of 2,814 and 1,249 tweets for train and test respectively covering five topics. These topics

²Scikit-learn <http://scikit-learn.org/>

are: Atheism (A), Climate Change (CC), Feminist Movement (FM), Hillary Clinton (HC), and Legalisation of Abortion (LA). As could be noticed, these topics are not just political, but actually covers topics of social (e.g. ‘FM’, ‘LA’) and religious (‘A’) domains. Chapter 3 section 3.4.1 provides additional details about this dataset.

We further used the Twitter REST API to collect the network information of the users in SemEval stance dataset. Basically, we collected two timelines for each of the users posted the tweets in our dataset, namely *Home* timeline³, which we use to construct the user’s IN; and the *Likes* timeline⁴, which we use to construct the user’s PN. In addition, we collected the user’s list of followers and friends to construct the user’s CN⁵. Unfortunately, we found that around 25% of these users have been deleted or suspended. Therefore, we end up with smaller number of tweets in the collection that we can apply our approach to them, exactly 1969 training and 906 testing data⁶. Table 6.1 shows the distribution of tweets (and users authored them) that we could retrieve in our dataset compared to the original SemEval dataset.

For each of the users in our collection, we managed to collect an average of 2,552 and 1,801 tweets from the *Home* and *Likes* timelines respectively. For each user, the set of mentions in those timelines were extracted and saved separately. In addition, we collected the set of friends and followers of each of the users in our collection. $IN_{@}$, $PN_{@}$, CN_{FR} , CN_{FL} represent the set of unique accounts appeared in the user’s tweets, likes timelines, list of friends, and list of followers of the user respectively. In addition, all the links appeared in the timeline were extracted and expanded (in case they were shortened). The domain of each link was then extracted and saved. IN_{DM} and PN_{DM} represent the set of unique web domains appeared in the user’s tweets and likes timelines respectively.

4.4.3 Baselines and Evaluation

We created two baseline systems that achieve the highest reported performance on the SemEval dataset based on the best performing participating system in the SemEval task (Mohammad et al., 2016b) that is trained on the *TXT* features. For the first base-

³https://developer.twitter.com/en/docs/tweets/timelines/api-reference/get-statuses-home_timeline.html

⁴<https://developer.twitter.com/en/docs/tweets/post-and-engage/api-reference/get-favorites-list.html>

⁵<https://developer.twitter.com/en/docs/accounts-and-users/follow-search-get-users/overview>

⁶List of ids of tweets and users network information would be made available

line, an SVM with linear kernel trained on the three stance classes using a combination of both word and character n-grams was used to represent the textual content of the tweet to be classified. Word n-grams was used with $n = \{1, 2, 3\}$, and character n-grams was used with $n = \{2, 3, 4, 5\}$. These features were used to train the SVM classifier with linear kernel. We only used the subset of training data that we managed to retrieve its users network information to allow a direct comparison to our models. The outcome of this model achieved an average F-score of 68.48 on our subset of the test data, which should be comparable to the reported best model in (Mohammad et al., 2016b) that achieved an average 68.98 F-score but on the whole dataset.

From a sociolinguistic perspective, it has been argued that there is no complete neutral stance as people use to position themselves with favor or against the object of evaluation (Jaffe, 2009). To comply with this argument, we created our second baseline by retraining the same SVM classifier with the same set of features, but with considering only the two polarised classes {favor, against} and neglecting the ‘none’ class. In this way, we force classifier to have a decision on the polarised stance of the user. While this approach will misclassify the samples in the test set with ground-truth ‘none’ stance, it was shown in the current state-of-the-art system (Siddiqua, Chy, and Aono, 2018) that this approach actually outperform the three-class classifier, where they achieved F-Score of 70% when trained a binary SVM classifier with tree kernel after neglecting the ‘none’ class. When we applied this approach, the overall F-score of the system got an actual improvement to reach 69.8%, which is comparable to (Siddiqua, Chy, and Aono, 2018).

After building the linear SVM baselines (both with the three and binary classes models), we trained the same models with the different set of suggested network features. We test each feature set separately and compare their performance to the models that depend on the tweet textual content; then we apply a different combination of the features to observe any potential improvement in the performance.

To evaluate the performance of our method, we used the official SemEval-2016 macro-average of the F1 score for the ‘Against’ and ‘Favour’, where the F-score on the ‘None’ class is discarded from calculating the average (Mohammad et al., 2016b). The same evaluation script provided by SemEval stance detection task was used to report the results. In addition, we show the performance over each of the five topics separately for a deeper analysis of the performance.

4.5 Results

4.5.1 Stance Detection Results

Tables 4.3 and 4.4 report the performance of the three-class classifier and binary classifier for stance detection, respectively. The general observation from the tables is that the binary classifier outperforms the classifier that is trained on three classes. While the binary classifier misclassifies tweets with no stance, it is more effective in detecting the polarised stance. This initial observation shows that forcing automatic classifiers to decide on a given stance might be a more effective approach than allowing them to have the ‘none’ option about stance, which makes it more confusing following Jaffe’s argument that there is no complete neutral stance (Jaffe, 2009). We analyse this further in the following subsection.

The second observation, for the binary classifier (Table 4.4), is that all the three set of network features - that are totally independent of the tweets contents - have better overall performance than the state of the art systems that depend on tweets textual content. In fact, the interactive network (*IN*) and the preference network (*PN*) features that combine the accounts and domains features achieve better results than the baseline on all the five topics. This confirms the consistent performance of network features over text on topics of different domains. In the connection network (*CN*), the friends network (CN_{FR} , the accounts the user follows) outperformed the baseline, while the follower network (CN_{FL}) achieved the lowest average F-score among all classifiers, even when combined with the friends network. This is potentially because of the sparsity of this network, where finding common followers among different users is less likely compared to finding common accounts they might follow, where it is expected to have people of similar stance following common accounts as a part of the homophily phenomena in social media (Al Zamal, Liu, and Ruths, 2012; Garimella and others, 2018).

While user’s interaction network showed the best overall performance among all feature sets, Table 4.4, it was interesting to see preference network outperformed all models in two of the five topics when using the binary classifier. These results support the hypothesis about stance detection, which is the online social network activity of a user posting a tweet contains enough signals to detect the stance of tweet regardless of its content. Furthermore, we show that the preference network of user’s likes on Twitter still can achieve decent detection of stance, which enables detecting stance for silent users.

Model	Topic					Overall		
	A	CC	HC	FM	LA	F_{favour}	$F_{against}$	F_{avg}
TXT (Baseline)	61.38	42.86	58.91	52.01	60.96	63.09	73.87	68.48
IN@	68.94	40.09	62.15	54.80	56.25	60.77	75.57	68.17
IN _{DM}	56.86	38.46	34.20	38.67	53.31	49.19	61.76	55.47
IN@+IN _{DM}	70.16	39.81	61.59	57.63	64.16	64.04	76.18	70.11
PN@	73.30	36.36	56.82	48.43	56.41	55.81	73.39	64.60
PN _{DM}	62.99	35.18	58.01	46.71	48.49	50.85	70.26	60.56
PN@+PN _{DM}	64.55	37.13	54.27	49.00	56.44	55.73	70.14	62.94
CN _{FR}	66.71	30.11	63.87	51.51	53.10	51.15	72.76	61.96
CN _{FL}	40.78	20.29	54.11	46.80	56.38	39.55	65.82	52.68
CN _{FR} +CN _{FL}	49.66	28.14	66.95	48.76	49.72	44.85	67.98	56.42

Table 4.3: Stance detection performance using different set of features using SVM classifier trained on three classes. F-Score (%) is reported on the SemEval stance detection task for each topic and overall. The set of features are categorized into three sets, namely, Interaction Network (IN), Preference Network (PN), and Connection Network (CN).

Model	Topic					Overall		
	A	CC	HC	FM	LA	F_{favour}	$F_{against}$	F_{avg}
TXT (Baseline)	61.91	42.86	59.53	52.21	62.40	63.53	76.07	69.80
IN@	68.30	54.14	59.05	50.40	60.82	61.89	77.90	69.89
IN _{DM}	63.24	42.86	53.91	61.24	60.17	61.51	76.82	69.17
IN@+IN _{DM}	67.65	42.86	62.64	55.87	63.93	64.04	79.07	71.56
PN@	73.49	42.86	59.26	49.63	63.87	63.70	77.70	70.7
PN _{DM}	67.14	42.17	58.33	51.62	61.79	60.18	77.28	68.73
PN@+PN _{DM}	68.03	42.86	59.00	52.57	65.50	63.91	78.60	71.25
CN _{FR}	63.83	42.86	64.01	60.93	59.58	64.53	78.25	71.39
CN _{FL}	35.97	42.86	58.51	52.70	62.68	56.08	69.73	62.91
CN _{FR} +CN _{FL}	50.00	42.86	68.21	57.38	54.13	58.07	73.41	65.74

Table 4.4: Stance detection performance using different set of features using *binary* SVM classifier. F-Score (%) is reported on the SemEval stance detection task for each topic and overall. The set of features are categorized into three sets, namely, Interaction Network (IN), Preference Network (PN), and Connection Network (CN).

Model	F_{favour}	$F_{against}$	F_{avg}
(A) TXT+IN@+IN _{DM}	67.21	76.49	71.85
(B) TXT+IN@+IN _{DM}	66.67	78.31	72.49

Table 4.5: The result of baseline linear SVM model when combining both text and network features. Model (A) and (B) shows the result when trained on three and two classes, respectively.

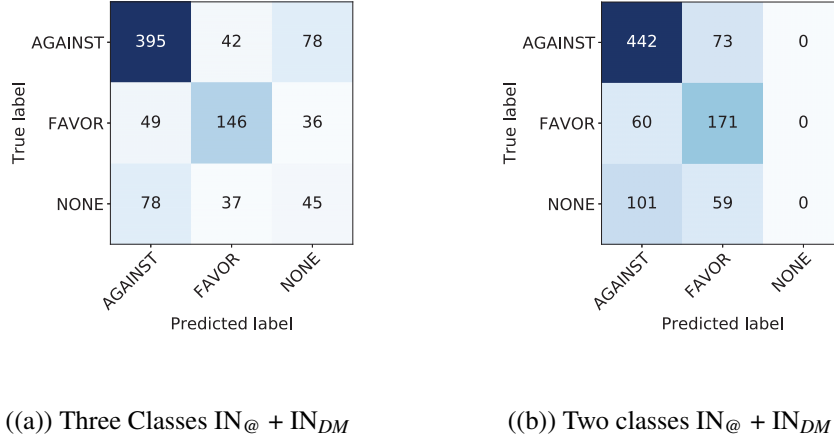


Figure 4.1: Confusion matrices for the best three vs two classes prediction models.

We further tested combining the best performing network features from the two networks (IN@+IN_{DM}) with (TXT) to see if this can further improve the performance. Table 4.5 shows the best achieved average F-score when we combined the network with content features, where the best performance achieved when we combined the interaction network with text for both the three-class and the binary classifiers⁷. This result was found to be statistically significantly better than the state-of-the-art baseline model using two-tailed t-test with $p - value < 0.05$ (we also tested significance using Mann-Whitney U test (McKnight and Najab, 2010), but it did not show significance).

4.5.2 Performance Discussion

As shown earlier, forcing the stance model to predict in-favor and against stances and ignore the ‘none’ stance consistently leads to better performance using all feature sets. This is an interesting result, since a binary classifier will always misclassify the ‘none’

⁷we also tested other combinations of feature sets, but TXT+IN@+IN_{DM} achieved the highest results

class leading to a larger number of false positives to the other two main polarised classes, which should reduce the performance. To better understand this, we plot the confusion matrices for the best performing model for both three/two class classifiers in Figure 4.1. As it is shown, the binary classifier led to a larger number of false positives for both the polarised classes; however at the same time, it led to larger number of true positives for both classes. This led to an improvement in recall with some reduction in precision, with an overall improvement in the average F-score.

Another observation from Tables 4.3 and 4.4, is the low performance of classifying stance on the features level. It can be noticed that using TXT performs better in LA (legislation of Abortion), while the performance on this topic changed when using a single subset of network features, mentions (@) and URL domains (DM). As using the combination of mentions and domains in interactions network (IN) provides better performance than using a single subset representation of features. Also, in TXT the features contain n-gram of characters and words of tweets, as provided by the best performing model in SemEval stance dataset (Mohammad et al., 2016b). Yet, the combination of connection network features (CN) lowers the performance of the stance detection in all the topics except Hilary Clinton. This might be due to the different representation these two interaction network carries on users level. Using follower (FL) features has the lowest performance in Atheism (A) in comparison with other topics. Moreover, it can be noticed that using a combination of friends and followers provides better performance in detecting the stance towards Hillary Clinton compared to other features. This kind of fluctuation is because of the clear polarization between political parties (political homophily in social relationships (Huber and Malhotra, 2017)), compared to the low polarization in the non-political topics. As the finding of (Huber and Malhotra, 2017) illustrates the effect of political polarization in forming a relationship as that people usually tend to form relationships based upon political similarity.

Also, we can notice the low performance of classifying stance on a topic level. As the climate change (CC) topic, where it has the lowest F-score among all topics. We conducted a further analysis and we noticed a large difference in the class distribution between the ‘in-favor’ and ‘against’ classes, where 176 samples in the training set are labeled as ‘in-favor’, while only 8 samples are labeled as ‘against’. This led the classification models to predict the majority class in most of the cases, which led to random-like performance for this topic.

Our obtained results for stance classification are the highest to be reported to date on the SemEval stance dataset, which confirms the large impact of utilising social

interactions demonstrated by user’s network activity as features in boosting the performance of stance detection, especially when combined with textual features. Our results highlight that user’s stance towards given topics could be inferred from various types of features from their activities online. In the following section, we apply an extensive analysis to these features to understand its role and influence in revealing the user’s stance.

4.6 Feature Analysis

In this section, we analyse each of the network features that showed to be effective in detecting stance. We apply our analysis to the binary classifier, which achieved the highest results. Our analysis includes studying the differences between our three networks, analysing most influential features per network and per topic, and giving examples of how these features might be effective.

4.6.1 Similarity between Networks

From the results obtained in Table 4.4, it is noticed that the scores achieved by the three groups of networks (IN), (PN) and (CN) are relatively similar. The average F-scores obtained by $(IN_{@} + IN_{DM})$, $(PN_{@} + PN_{DM})$ and (CN_{FR}) are around 71% and their results were found to be statistically indistinguishable from each other using both t-test and Mann-Whitney U test. This motivates to further examine the overlap among these networks, since it is highly possible that users interact with and like content of the same set of accounts they follow. Hence, we measure the overlap between the features of (IN), (PN) and (CN) to gauge the similarity among them.

For each user, we compute the similarity between their $IN_{@}$, $PN_{@}$, and CN_{FR} features using Jaccard similarity, then we plot the distribution of the similarity score across all users. We repeat this process for the domains features by computing the similarity between IN_{DM} , PN_{DM} . Figure 4.2 shows the similarity distribution between the network’s sets, where zero indicates no overlap and 100% means identical sets. We observe that there is a noticeable difference in each network for the same feature component. The overall similarity between accounts in each of the three networks ranges between zero and 20%, and it ranges between 0 and 35% for domains. This result means that users tend to interact and like contents from users out side their connection network, and like tweets with links generally different from the domains they link in

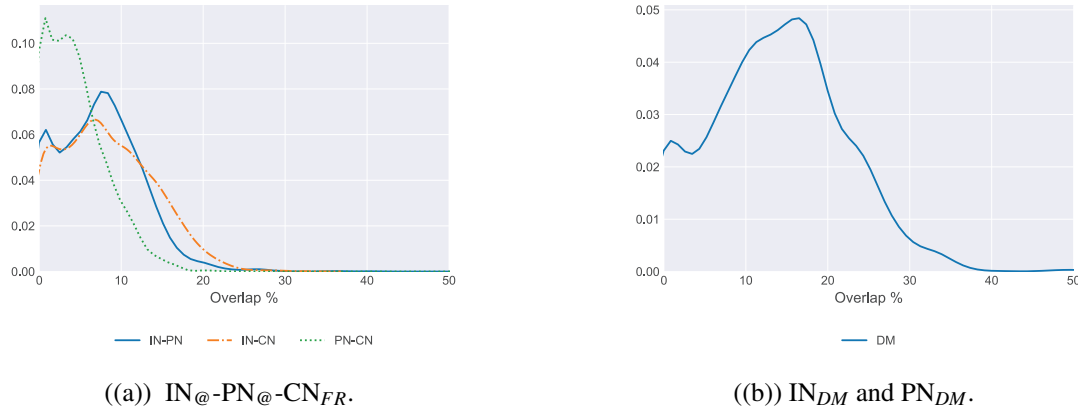


Figure 4.2: Similarity between CN, IN and DM in users dataset.

their tweets. This is actually an interesting finding, which actually raises further research questions about the reason of having the performance of the three networks in stance detection similar when they are mostly different.

There is a passable explanation behind the similar performance, that actually the small percentage of similar accounts (domains) between the three networks are those which create the most influential features for the classification, and thus the three classifiers achieved comparable performance. Therefore, we further analyse the similarity between the most influential features of the three networks sets, where influential features are identified as those having the highest weights for each of the classes for each topic. We use Jaccard similarity to compute the similarity between the top N influential features of IN, PN and CN and plot the similarity for $N=\{1 \rightarrow 1000\}$. Figure 4.3 presents the similarity for each network features influencing favor and against stance. Again, it is observed that similarity between the most influential features is not high for any of the networks for both ‘favor’ and ‘against’ classes, where the similarity does not exceed 10% for the accounts, and 17% for domains.

These findings confirm the differences between the three networks, and show that each network represents a different set of accounts and domains for the same user. Even with the most influential features for models trained on each classifier the set of features is different than the other. This means that a more in-depth analysis to these features is required to understand the high performance of the classifiers trained on the three fairly independent networks.

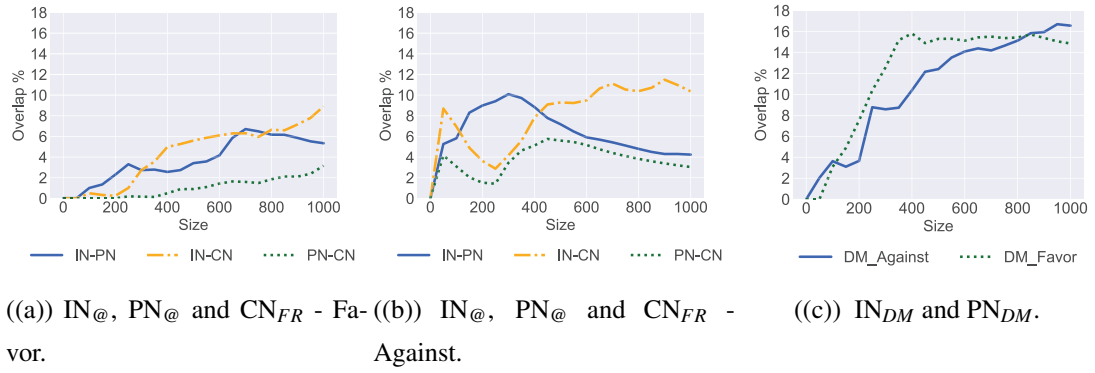


Figure 4.3: Similarity between CN, IN and DM for (In-favor and Against) stances with respect to the top features.

T	NW	Favor	Against
A	IN	@atheism_tweets, @atheistrepublic, @god_stupid	@ChristianInst, @godlesstheory, @godbiblechurch
	PN	@thetweetofgod, @foxnews, @nytimes	@prayerbullets, @reuters, @cnn
	CN	@Stephenfry, @RichardDawkins, @MarilynManson	@baptism_saves, @srisri, @artoffiving
CC	IN	@telegraph, @independent, @climareality	@skynewsbreak, @nytopinion, @reuters
	PN	@nytstyles, @news4anthros, @fox2now	@cnn, @foxnews, @nyhealth
	CN	@barackobama, @potus, @mashable	@foxnews, @sentdcruz, @cnn
HC	IN	@washtimes, @hillaryclinton, @realdonaldtrump	@trumpstudents, @foxnewssunday, @brianschatz
	PN	@cbsnews, @nbcnews, @hillaryforh	@govchristie, @drbiden, @sentedcruz
	CN	@hillaryclinton, @billclinton, @shehasmyvote	@foxnews, @realdonaldtrump, @madam_presiden
FM	IN	@mtv, @goodreads, @feministculture	@feministfailure, @goodmenproject, @womenwriters
	PN	@feministajones, @ppfa, @foxnews	@nytopinion, @mtvnews, @weneedfeminism
	CN	@vday, @Schofe, @twitterfashion	@Truth_seeeker, @femalefedupwith, @thepowhouse
LA	IN	@humanesociety, @skynews, @ppactionca	@ppfa, @nbcnews, @bible_time
	PN	@savewomenslives, @dallasnews, @citynews	@onejesusloves, @younglife, @yahoonews
	CN	@thedemocrats, @barackobama, @hillaryclinton	@prolifeyouth, @march_for_life, @lifeteen

Table 4.6: Top features extracted from the best model in each case and trained on two classes, CN_{FR}, IN@, PN@.

4.6.2 Which Network Features Reveal the Stance?

To get meaningful insights about the contribution of the features to infer the stance, we identify the most influential feature of the best model from (CN), (IN), and (PN) network with regards each topic. We hope this would give some explanation to the good performance of these models, especially after we found that these networks do not highly overlap.

Table 4.6 and 4.7 show the top features that have a noticeable influence on the stance classification for each topic with respect to the weights of features in the linear SVM model for the best features from each network group: (IN@+IN_{DM}), (PN@+PN_{DM}) and (CN_{FR}).

In the (CN_{FR}) network, the social influence manifest through the users' friends (following network). Users tend to follow the accounts that support their stance. For

T	DM	Favor	Against
	IN	sciencealert, thinkprogress, washingtonpost	nationalpost, washingtonpost, newsweek
A	PN	reuters, newhumanist, telegraph	faithreel, bible,prayerbullets
	IN	thetimes, nytimesarts, nbcnews	bbc, naturalnews, washingtontimes
CC	PN	abc, newswire, nypost	cbc, telegraph, washingtontimes
	IN	nytimes, thedailybeast, cnbc	opposingviews, washingtontimes, foxnews
HC	PN	nytimes, theguardian, nbc	cnn, foxnews, newsfoxes
	IN	cnn, buzzfeed, nytimes	dailymail, bbc, theguardian
FM	PN	apnews, washingtontimes, feministing	independent, dailymail,activistpost
	IN	newstatesman, nytimes, cnn	nypost, dailymail, cbsnews
LA	PN	bpas, ahealthblog, thenation	lifeneews, gotquestions, cnsnews

Table 4.7: Top features extracted from the best model in each case and trained on two classes, IN_{DM}, PN_{DM} .

instance, users with against stance toward legalisation of abortion (LA) tend to follow accounts that oppose the abortions such: '@prolifeyouth', '@march_for_life'. The same for the users with favor stance to Hillary Clinton where the top followers are '@Hillaryclinton', '@billclinton', '@shemyvote'. Users who have a favor stance towards Atheism tend to follow social actors with the same believes such: '@Stephenfry', '@RichardDawkins', '@MarilynManson'. Similarly, users with favor stance toward feminist movement follow the accounts that support feminism. One of the top features that identifies the in-favor stance toward feminism is '@vday', which is an activist movement account that supports the feminist movement as this account description indicates: "to End Violence Against Women Girls.". For the climate change and legislation of abortion, the politicians and news outlets are the most influential accounts in predicting the stance. We can not specify whether these users follow such account because they support their opinion towards each topic.

Unlike CN, influential accounts for IN and PN include news accounts. For instance, the news accounts '@washtimes' and '@cbsnews' are one of the distinguishing features to detect the favor stance to Hillary Clinton in $IN_{@}$ and $PN_{@}$. In addition, '@telegraph' in $IN_{@}$ has a positive correlation with favor stance to climate change. Users with favor stance to the legalization of abortion interact with '@skynews' account. In contrast, news accounts have a minimal effect in detecting stance toward feminist movement and atheism, where the top mentions features that capture a favor stance are accounts that support the topic: '@atheism_tweets' and '@feministcultur'.

Also, another difference between IN and PN, is that IN usually contains accounts

of opposing view since in this case the interaction can be through replying or quoted retweets with opposing comments. This case can be seen in Hilary Clinton topic, where '@realDonaldTrump' is one of the top features for the 'favor' stance in IN. It can be imagined that the interaction here is not for support as shown in table 6.4 (Example 3). In addition, interacting with accounts that have a related meaning to the topic seems to have a visible correlation with detecting the against stance of users. For instance, the interaction with '@godlesstheory' and '@godbiblechurch' has an influence in detecting the against viewpoint to atheism. Similarly, '@bible_time' captures the against stance toward abortion. Furthermore, famous accounts with clear support to a related social issue have a clear influence in detecting the stance. For instance, users with against stance to feminism interact with '@feministfailure'. In addition, users who oppose the legalisation of abortion interact with '@ppfa', Planned Parenthood account.

For the web domain features, it can be noticed that the top domains features IN_{DM} and PN_{DM} are mostly news websites. News websites and media outlets such as 'washingtonpost' and 'sciencealert' are one of the distinguishing features to detect favor stance toward Atheism. In contrast to mentions, the news websites have a noticeable effect in detecting users view points toward feminist movements. We can see that users with against stance to feminist movement tend to share contents from 'dailymail', 'bbc' and 'theguardian' websites. Users with support stance to feminist movement tend to share contents from 'cnn'. Users with against stance to Hillary Clinton share contents from news websites such as 'opposingviews', 'washintontimes' and 'foxnews'. The website 'nytimes' has a positive effect in identifying the favor stance to Hillary Clinton. We can notice some overlap between IN_{DM} and PN_{DM} , where it seems users like and interact with the same news and media outlets in the PN and IN networks. For instance, users with against stance to Hilary Clinton tweet interact and like news contents from 'foxnews'. The same for users with against stance to the feminist movement, the users like and interact with 'daily-mail'. In general, there is a tendency for the users to like and share content from the same media as described in the next section.

4.6.3 The Context of the Features

We carried a further qualitative analysis to identify the context in which the IN and PN features correlate with the topic of the target. Table 6.4 shows a sample of tweets from the users' timelines (IN) and Favorite timeline (PN) with respect to topic-stance pair

#	T	Feat	Example tweets (favor)
1	CC	IN@	RT @Telegraph: Prince Charles reveals his gardening inspiration: a hidden Buckingham Palace veg plot https://t.co/tBZB5DSKt5
2	HC	PN@	@NBCNews Kill the bear for BEING A BEAR! What's wrong with this?
3	HC	IN@	You are an idiot on so many levels, @realDonaldTrump https://t.co/keptgYgTed
4	FM	IN@	I'm your nightmare come true," said Angela. #YAlit #vampire #paranormal #Action #humor https://t.co/MCvYEvdz8Z @goodreads
5	A	IN@	@god_stupid @userid just the ignorant, racist, sexist, child abusing fanboys that roll play #christianity.#Atheist and proud
#	T	Feat	Example tweets (against)
6	CC	IN@	RT @SkyNewsBreak: Former Labour Prime Minister Tony Blair has told Sky News Theresa May will win the General Election #GE2017
7	A	IN@	RT @ChristianInst: Romans 8:28 And we know that for those who love God all things work together for good, for those who are called accordi
8	A	PN@	@prayerbullets: Turn every curse sent my way into a blessing -Neh. 13:2 #Prayer

Table 4.8: Sample of tweets and the context of IN and PN in relation with stance and topic.

and highlights the social interactions with the top features. As explained in the previous section, what sets apart users with support/against stance to climate change are those pertaining to news portals. For instance, the most dominant mentioned accounts that influence the supporting and opposing position toward climate change is '@telegraph' and '@SkyNewsBreak'. Users interaction with these news accounts in the sense of re-tweeting and liking the news that has no relation to climate change (Example 1 and 6).

tweets from '@NBCNews' with no relevance to Hilary Clinton or the presidential candidates tend to be liked by users with a stance supporting Hillary, (Example 2). The same with users who support feminist movement, they interact with account '@goodreads' with no topical relation to stance topic, (Example 4). The user mentioned @goodreads to promote to the novel "Beginnings" which is a teen romance, sci-fi and fantasy story. Furthermore, example 6 demonstrates how the interaction with @SkyNews helps in predicting the against stance towards Climate Change (CC) even with news that does not concern with climate change. In contrast, users opposing atheism tend to mention religious accounts to support their stance against atheism. For instance, users with against stance toward atheism interacted with '@Christian-

Inst' by retweeting verses from scripture (Example 7). Furthermore, users who have an against stance toward atheism tend to like religious's content from accounts such as '@prayerbullets' (Example 8). Users supporting atheism interact with accounts that are sarcastic toward religions such as '@god_stupid' account, in a sense of hashtag as a way of expressing the against viewpoint towards the religious people. The account '@god_stupid' is a sarcastic account, yet the interaction with it tends to take a kind of attacking the religious means as shown in (Example 5). Similarly, Users supporting Hilary Clinton defending their viewpoint by attacking '@realdonaldtrump' (Example 3).

4.7 Discussion

In this chapter, we studied the possible features for multi-modeling the stance on social media and how these features may reveal the user's stance from their publicly available online data. Unlike most of the literature in this area, which mostly focuses on achieving high accuracy without in-depth analysis, our main focus is to understand how stance could be revealed throughout different sets of signals. This led us to explore multiple sets of features including some that have not been examined before (such as the preference network), and test it on a stance benchmark dataset of multiple topics of different genres.

4.7.1 What factors predict the stance?

Our study in this chapter investigates three research questions that have not been sufficiently explored in earlier studies on stance detection.

Our first research question "*RQ2.1*" is concerned with exploring the different signals from user's public social media profiles that can reveal their stance. We have defined three sets of network features, including interaction (IN), preference (PN), and connection (CN) networks, and compared their performance to textual features that represent the state-of-the-art models on the SemEval dataset. Our findings showed that user's stance can be detected with many signals, including textual content and different sets of network features. We found that using network features leads to a more accurate stance detection than using content-based features solely, and the performance becomes statistically significantly better when both sets of features are combined together. We also noticed that when building a stance classifier, a binary classifier is more

superior than a classifier that allows neutral stance, which could be linked to the argument that there is no “neutral” stance and everyone should have some leanings (Jaffe, 2009).

Our second research question “RQ2.2” focused on how the performance of the stance detection using these features would differ across different topics. Our analysis of the five topics in our dataset showed that network features consistently achieve better performance on average compared to textual features. We only found that the performance for one topic (CC) has always the lowest F score. Our investigation to the distribution of the stances on this topic suggests that the problem stems from the large imbalance in the training samples, which leads the prediction model to predict only the majority stance class, which is independent of the set of features used.

As for our third research question “RQ2.3”, which concerns with investigating what makes the introduced features effective for stance detection; we initially analyzed the overlap between the accounts and web domains for each of the users in our dataset in the three networks: IN, PN and CN to ensure that their similar performance is not the reason for their high similarity in their nodes. It was surprising to find them mostly dissimilar with low overlap between them with $\leq 20\%$ similarity between them. This was interesting to see that each of them captures one side of the user’s activity, and each can reveal their stance. We further investigated the top features in each network model. We noticed that the top features can sometimes be topically unrelated to the target and yet have a high impact on deciding the stance of the topic. For instance, the interactions with accounts as @goodreads and @SkyNews help in detecting the stance towards feminist movement (FM) and climate change (CC) respectively, as shown in section 4.6.3. Since these features have no direct relation to the topic of the stance, this indicates that the user’s stance can be detected with many signals regardless of the topic. We showed that using content-less features help in detecting the stance for the users with an implicit point of view toward a topic where the users may not directly express their point of view by using keywords related to the target. As the top features extracted from the two networks (PN@) and (IN@) have no direct relation to the stance’s topic. For instance, the ‘@Telegraph’ was one of the top features that predicts the in-Favor stance towards Climate Change (CC) topic.

Furthermore, one of the key findings from this study is the high performance of PN and CN for stance detection, which outperforms the state-of-the-art baseline TXT model. This shows that detecting stance for silent/passive users (who never tweet or share any content) is doable, given the condition that they have enough common sig-

nals in their preferences and connection networks. This raises a real concern about the privacy of social media users in general, and motivates future research in the direction of protecting those users from having their leanings and beliefs revealed unconsciously (Waniek et al., 2018).

4.8 Summary

To answer the second research question of this thesis "RQ2", we illustrated the multimodality of stance detection as motivated by the sociolinguistic theory by (Bassiouney, 2015), where it defines stance as the link between linguistic forms and social identity. The experiments and analysis were applied to a set of five political, social, and religious topics.

The finding of this chapter suggests that using network features leads to a more accurate stance representation than using content-based features solely. Using contentless features helps in detecting the stance of users with an implicit point of view toward a topic where the users directly express their point of view by using a keyword related to the target. We used to detect the stance leverages the network features from two streams: the timeline stream and the favourite (likes) stream. Deriving network features from user direct interaction managed to enhance the performance of the overall stance detection model. These kinds of features have no topical relation to the target, and yet the stance representation improved for the given target. Yet, users tend to interact with accounts in the mean of expressing their support/opposing view regardless of the account contents.

Another finding shows the possibility to predict the stance of the passive users who do not have direct interactions through their preference network. This finding illustrates that regardless of the topic, there are usually common signals in the users' activity and preference networks that can indicate the stance of those users towards this topic. Next, we describe a framework to obfuscate social media users' stances.

Chapter 5

Stance obfuscation

Chapter 4 shows that the unconscious online social signals can reveal the viewpoints towards a topic. The amount of social data generated by social media users is unprecedented. This makes the social media platform a valuable source to collect social data and increase reliance on algorithmic predictions. These prediction algorithms, which leverage users' online data, have intensified the call for a mechanism to preserve the privacy of social media users. There has been a noticeable controversy around using traceable digital data without having the users' consent. In this chapter, we seek to answer the third research question, *RQ3: What is the minimal number of online signals that a user can inject or remove from their social media activity that can mislead stance models from predicting their stance?*. To answer this question, we use two methods to produce a divergent on the prediction models' overall performance. However, the following question remains: Can a normal social media user decipher the relation between an online signal and the prediction algorithm result? Here we show the complexity of interpreting the relation between the result of the prediction model and the online signals. Hence, human intuition cannot be used as a reliable guide to obfuscate their online social identity.

5.1 Introduction

In the realm of social media, considerable personal data can be collected for each social actor. In this online environment, a user's attitudes can be easily predicted from the online social signals without them having to explicitly express such information. However, there is substantial dependency of the social researcher on the social sensing data to predict people's attitudes and social behavioural patterns. For instance, users

may wish to conceal aspects of their identity to avoid harassment (Reddy and Knight, 2016). Social media platforms cast a special type of complexity to preserve the identity and sensitive information as such platforms are expected to be open to allow social interaction. Balancing the trade-off between openness and preserving user identity is a hard dilemma.

This urges the need to reinvent a societal technique that can help the social actors to have a choice in opting out and preserving their sensitive information from being predicted. Previous studies have shown the need for an ethical framework to protect user identity (Reddy and Knight, 2016; Fiesler and Proferes, 2018; Williams, Burnap, and Sloan, 2017). However, there have been limited contributions in this domain. The authors in (Reddy and Knight, 2016) introduced a framework to obfuscate gender identity from textual data and generate a divergent text that is fluent and semantically similar to the original text. In addition, the authors in (Perez, Musolesi, and Stringhini, 2018) proposed a metadata randomization of online social media users. Their method depended on the integer representation of users' metadata on social media. However, this method could not sufficiently protect and obfuscate user identity. Another study by (Waniek et al., 2018) proposed a theoretical framework focusing on the network representation. In contrast, the overarching aim of this paper is to propose a framework to balance the needs between the social researcher and the users' usage of social media: How to balance the needs of using social sensing data and preserving the privacy of the users' usage of social media?

This framework promotes an ethical practice for social scientists who use social sensing data in a decision-making algorithm. This framework is based on analysing a machine learning algorithm to decompose the features that reveal one's non-expressed information. Furthermore, through a survey study, we further highlight the inability of social media users to decipher the online social signals which support the main intuition behind the requirements of the proposed framework. In particular, this chapter studies the construction of an obfuscation framework concerning the optimality of this method in relation to the user needs. Furthermore, we are guided by a real user's perceptions on their information being disclosed on the social media platforms. This study aims to design an open algorithm to share the data with a user consent control (Hardjono, Shrier, and Pentland, 2016). In contrast to the literature, this study mainly focuses on stances expressed on the social media and developing an obfuscation method to obscure the users' viewpoints and learnings. In particular, this chapter studies the third research question [**RQ 3**: What is the optimal number of online sig-

nals that can be added to/removed to degrade the performance of the stance detection model? We develop a stance obfuscation framework on social media, with an objective to facilitate the design of an open algorithm for data-sharing with user consent control (Hardjono, Shrier, and Pentland, 2016). Social media platforms have a complex nature of preserving the identity and sensitive information of its users. Such platforms should be open to allow social interaction. Balancing the trade-off between openness and preserving privacy is a hard dilemma. We design a stance obfuscation strategy with a mixed set of content and network signals. To address the third research question, this chapter introduces a framework for the obfuscation strategy, which is embedded with a practical development of methodological actions to affect the performance of the stance detection model. This method is evaluated using the topics derived from the SemEval dataset. Furthermore, we analyse the most influential factors in obscuring the user's stance with respect to various topics.

5.2 Related work

In this section, we discuss the studies that have used obfuscation techniques to preserve the identity of social media users. These studies can be categorized based on the type of online signals that can reveal users' privacy, including textual content and interactions on social media.

The previous studies that have used textual content to obfuscate the identity of a social actor have mostly been concerned with preserving the semantic of the textual content while providing the required divergent of the social actor's identity. For instance, the authors in (Reddy and Knight, 2016) introduced an obfuscation framework, using lexical substitution, of Twitter and Yelp posts to preserve gender information and confound a demographic classifier to predict the opposite gender. These authors used a subsided method where an input text W was transferred to a new text W' while preserving the original meaning. This method mainly depends on the *word2vec* extension of Levy and Goldberg (2014) to execute the lexical similarity method and generate similar candidates for a given token. In addition, the authors in (Yang, Qu, and Cudré-Mauroux, 2019) proposed a user meta-data preserving method for ranking social media content. They used a special type of differential privacy exponential mechanism where the algorithm uses a probability function that decreases exponentially with the distance between the old and new proposed obfuscating content. The authors in (Xu et al., 2019) introduced a privacy-aware text rewriting method on behalf of dataset providers

instead of a social actor by rewriting the text. They used three datasets representing three types of identities: race (Blodgett, Green, and O'Connor, 2016), political affiliation (Voigt et al., 2018) and gender (Reddy and Knight, 2016). The work of the introduced methods has been evaluated based on two criteria: a) linguistic quality of the sentences by evaluating the semantic of the text and b) obfuscation of the sensitive attribute to reduce the leakage of sensitive information by reducing the overall accuracy of the identity prediction model. They showed that the fair risk method provides a robust result in comparison with adversarial training methods.

The other line of work concerns preserving the social media identity using non-textual signals, such as users' meta-data. For instance, the authors in (Perez, Musolesi, and Stringhini, 2018) introduced a user identification obfuscation framework by randomizing the users' meta-data to preserve their online identity. These authors defined identification as a classification problem to build behavioral signatures for each user. In addition, the authors in (Jakaza, 2020) used three methods—lexical, textual pragmatic and interactional levels—to design an obfuscation framework for Facebook and Whats-App. They analyzed the effectiveness of the obfuscation methods from a socio-semantic perspective and used a randomization method among a set of three features.

In contrast, we focus on one aspect of identity, '*stance*', by leveraging online interactions to design a framework for stance obfuscation. The formulation of this research problem can be considered as defined by Reddy and Knight (2016), with an objective to degrade the performance of the stance detection model. The next section explains the aspect of this framework and the experimental design used to evaluate the proposed obfuscation methods.

5.3 Obfuscation framework

The stance on social media can be modelled using different online features. Chapter 4 provides a comparative analysis of two types of features: interactions and content-based features. As the study in the previous chapter demonstrated that the interactive network (IN) and connection network (CN) enhance the prediction score of the stance model, in this study, we use these two sets of features as the input for the obfuscation framework.

5.3.1 Stance obfuscation feature space

This framework aims to generate divergent cues that can degrade the performance of the stance detection model and affect its robustness. The proposed algorithm takes as input Y as the stance label along with a set of features X , which represents the features derived from IN and CN , as demonstrated in Chapter 4. We use two methods to transform X to a new set: a) data encapsulation (DE) and b) data removal (DR). Our transformation search space is simple: each feature in X can be substituted with another feature, as guided by the best performing stance detection model, to generate a substitute feature set X' . The next section explains the feature space of the stance detection model X , along with the two obfuscation methods.

For each user in the dataset with stance labels towards any of the five topics (feminism, Hilary Clinton, atheism, climate change and abortion) is specified as either ‘in favor’ or ‘against’, we use two obfuscation methods to substitute the feature vector associated with that user according to the top influential features T in predicting the opposite stance by using the coefficients of the best performing stance model, as explained in Chapter 4.

5.3.2 Obfuscation methods

We use two methods to create a divergent for each user’s stance in the dataset, namely DE and DR . We formalize the methods as an optimization problem whose objective is to minimize the performance of the stance model by modifying the data.

DE: In this method, we use the top features T to add additional values to each user feature vector. For the two stance labels ‘favor’ and ‘against’, we add opposite features to the user’s stance. $T_s = \{t_1, \dots, t_i\}$, where t_i corresponds to the value of X_n of the top predictive features for the stance $s = \{\text{favor, against}\}$. The value of the most predictive features for each class T is extracted using the best performing model, as explained in Chapter 4. The below equation demonstrates the basic functionality of the ‘encapsulation’ process, where T'_s is the set of top features obtained from the opposite stance to the user with feature set X_s .

$$E(X_s, T'_s) = \{x_1, \dots, x_n, t_1, \dots, t_i\} \quad (5.1)$$

DR: In this method, the top features T are removed from the user feature vector to create a divergent on the prediction model. For each user feature vector x , the method

removes the corresponding set of top features T_s , which can be represented as $X_s - T_s$, where s demonstrates a similar stance to the users with feature set X_s .

5.4 Experimental Setup

5.4.1 Dataset

We assume that the stance detection algorithms have access to a training dataset. We continue using the well-known stance benchmark dataset ‘*SemEval 2016 stance dataset*’, as explained in Chapter 3, and remove instances that do not have a clear polarized stance, namely instances with ‘*neither*’ stance. This dataset contains five topics (atheism, climate change, feminism, Hillary Clinton and abortion), and the stance is inferred by a given tweet as being in favor, against or neither. We keep only the tweets that satisfy the following conditions: (i) they indicate the user’s stance as either ‘*in-favor*’ or ‘*against*’ and (ii) their authors did not have their account deleted or suspended at the time of our study. Table 5.1 illustrates the distribution of tweets/users in the dataset.

Topic	Users	Home timeline tweets	Accounts (interactions)	Domains (interactions)	Friends (contacts)
Atheism	550	832773	121548	5834	375695
Climate change	461	607095	135915	8938	280536
Feminism	524	778512	165107	6384	250537
Hillary Clinton	670	1043282	208426	7111	435612
Abortion	670	924734	190848	7725	405269
Total	2875	4186396	821844	35992	1747649
Unique total	2234	4028574	721354	25647	1365075

Table 5.1: Distribution of tweets in the dataset for the five topics.

5.4.2 Stance detection models

To evaluate the effectiveness of the proposed obfuscation methods, we use four machine learning models for stance detection and train them on the substitute feature set $SubstF$. To extend the experiment to different types of stance detection, we use the following models: support vector machine (SVM), logistic regression (LR), convolutional

neural network (CNN) and naive Bayes (NB). For LR, we use "LGBTQ" with the random state set to zero. For CNN, we use three layers and compile the network using the Adam optimizer. Finally, to implement the NB algorithm, we use the multinomial NB as the configuration for the stance detection model. Table 5.2 shows the macro F1 score of the models trained on the SemEval stance dataset. The above-mentioned classical machine learning algorithms are found to outperform the more advanced deep learning alternative, CNN; these results align with the previous findings obtained from 2 on evaluating machine learning algorithms using the SemEval stance dataset.

Prediction Model	Contact			Interactions		
	In favor	Against	Macro F1	In favor	Against	Macro F1
Random baseline	51.16	74.18	62.67	51.16	74.18	62.67
SVM	71.55	86.59	79.07	72.00	86.92	79.46
LR	73.42	87.62	80.52	71.30	87.21	79.26
NB	70.82	82.37	76.59	76.66	87.25	81.96
CNN	62.25	88.56	75.40	62.36	84.61	73.48

Table 5.2: F1 scores of stance detection algorithm before the hiding process.

5.5 Results and discussion

5.5.1 Effectiveness of obfuscation methods

We use the macro F1-score to validate the effectiveness of the proposed method, as advised by previous studies (Reddy and Knight, 2016). The lower the overall score generated by the stance detection model, the better is the obfuscation strategy. Figure 5.1 illustrates the results of the two techniques, *DE* and *DR*. Overall, the *DE* method has a significant effect on degrading the performances of the four stance detection models. While the *DR* method has a relatively minimal effect, decimal decrees in the performance of stance detection models. The *DE* method results in better obfuscation as the resulting substitution feature space X' contains more features that are not correlated with the target stance label. In contrast, the *DR* method has a shortcoming as the proposed obfuscation feature set (T_s) might not be present in the original feature space X of the user.

Another observation on the level of models is that the proposed obfuscation methods are effective for the classical ML models of stance detection. This is attributable to

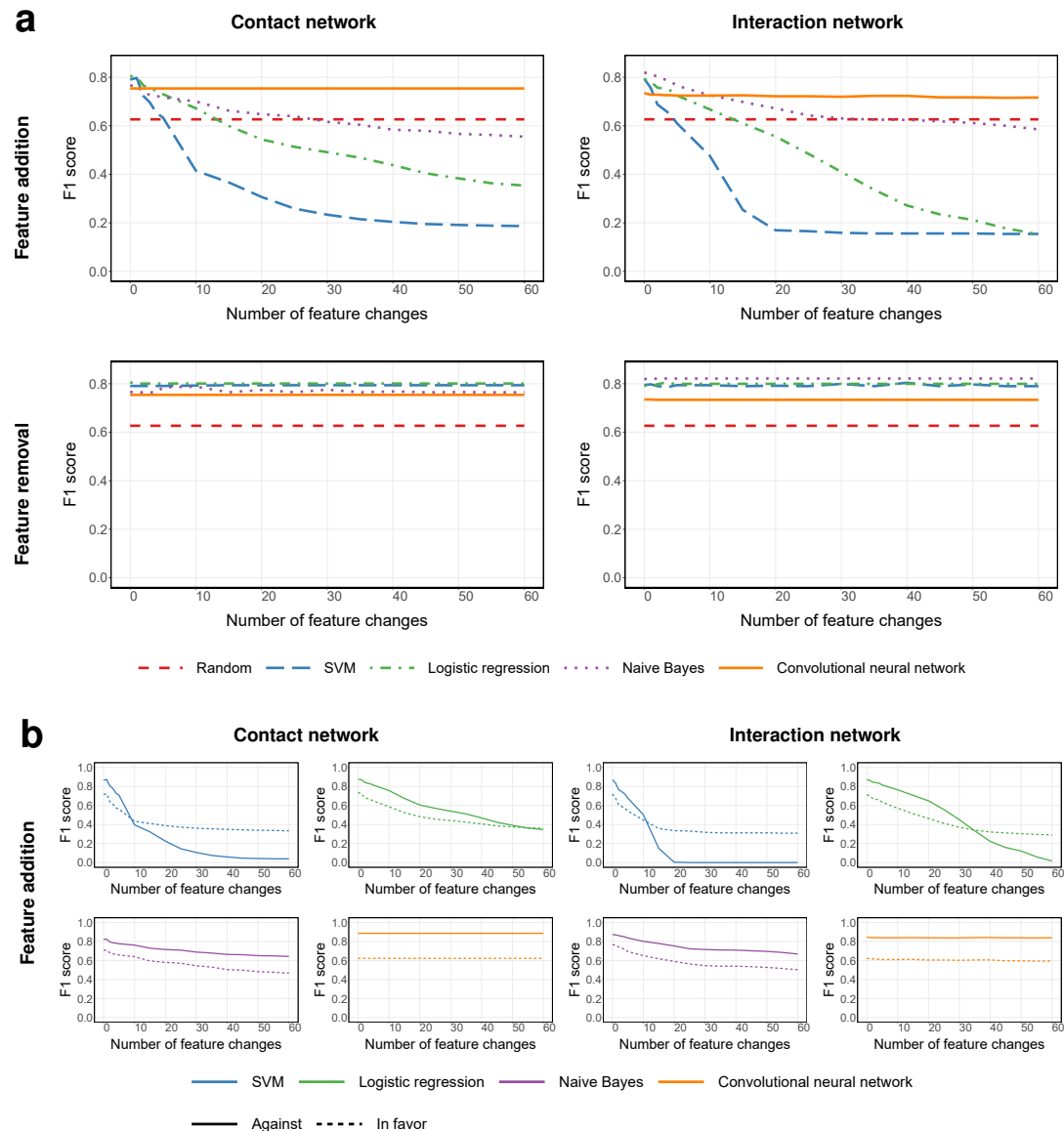


Figure 5.1: Effectiveness of stance obfuscation on four stance detection models—SVM, LR, NB and CNN, in comparison with the random stance detection model.

the fact that DE provides an effective stance divergent after adding an approximate of four contact/interaction features. The effect on these models is noticeable as the overall performance is comparable to that of a random classification model. In contrast, the obfuscation methods are the least effective on CNN in comparison to the other stance detection models. The DE method provides a decimal decrease in the overall CNN performance. Overall, while CNN is not the most effective approach for stance detection yet there is a possibility to further enhance the obfuscation methods to tackle the deeplearning algorithms. Moreover, it shows that the divergent of the stance prediction result can be achieved by adding the average of the four values in the substitute feature

set X' to disrupt the prediction model and reach a random prediction behaviour.

5.5.2 Additional paradigm

To provide an additional motivation to the need for the proposed framework from the perspective of social media users, we conduct a survey study to evaluate the ability of social media users to anticipate the influenceability of the social signals on the prediction algorithm. The study surveys approximately 1,000 participants recruited through Amazon Mechanical Turk. The main objective of this study is to evaluate the users' ability to decipher the stance using online social signals as cues. The participants are provided a set of social online signals related to three topics: Atheism, Hillary Clinton and Feminism. First, we examine the degree to which Twitter users feel the need to avoid revealing their stance. It is found that participants with strong stance towards a topic, strongly against or strongly in favor, are less inclined to reveal their true stance on Twitter. This can be noticed in Appendix B.1 figure B.2, which presents the proportion of participants who report a value of 6 or above when assessing their need to avoid revealing their stance on Twitter.

The proportion of participants who report 6 or above ranges 14% to 28% for those with a strong stance and 30% to 45% for the remaining participants, which indicates that the need to unreveal a stance is a result of stance uncertainty rather than the need to unreveal the stance.

In addition, we evaluate the participants' ability to decipher the predictable features of stance online. For each topic and each feature type (word used by the user in a tweet, account followed by the user and account mentioned by the user in a tweet), we select three features that are most indicative of being 'against' the topic according to the stance classifier, as well as three features that are most indicative of being 'in favor' of the topic. The findings show that for all feature types, the participants are less capable of identifying the features associated with the 'against' stance than those associated with the 'in-favor' stance. At the topic level, as Figure B.1 in Appendix B.1 shows, the overall percentage of participants who correctly identify the features is less than 50%. These findings illustrate that people cannot rely on their intuition to recognize the features that reveal their stance on social media.

5.6 Summary

This chapter seeks to answer the third research question, with an objective to balance the needs of using social sensing data and preserving the privacy of the user's social media? To address this question, we evaluate the social media users' ability to decipher their identity represented by stance. The findings illustrate the complexity of normal users to interpret the social signals and their relation to the general attitude that can be inferred using the prediction models. Evidently, the people's intuition about the correlation between the social signals and attitude towards a topic is not correct. The experimental results demonstrate the effectiveness of the encapsulation method in obfuscating the stance in comparison with data removal. While CNN is not the most effective approach for stance detection compared to other simpler methods, it is the most robust against the obfuscation methods.

In the next chapter, we assess the finding derived from Chapter 4 and analyse the most effective features in predicting the stance to examine the interplay between stance and different types of accounts.

Chapter 6

Characterizing the role of bots' in polarized stance on social media

In Chapter 4 we show a comprehensive analysis of the online factors that affect stance detection on social media. The findings brought the attention to other contributing online factors that have an important effect on stance detection. In this chapter, we seek to answer the fourth research question "*RQ4: How the stance detection model can be used to evaluate the interplay of bots with online stance?*". As demonstrated by the main finding of Chapter 4, the interaction with other users is shown to be one of the influential features in detecting the online stance on social media. Moreover, we show in Chapter 2, that most of the previous work on stance detection have been conducted to analyze the stance towards an event/topic. In this chapter, we extend the effort to utilize stance detection to analyze the effect of automated accounts on online stances.

There is a rising concern with social bots that imitate humans and manipulate opinions on social media. How do interactions with social bots may play a role in affecting online users' stances on a given topic? Current studies on assessing the overall effect of bots on social media users mainly focus on evaluating the diffusion of discussions on social networks by bots. Yet, these studies do not confirm the relationship between bots and users' stances. This study fills in the gap by analyzing if these bots are part of the signals that formulated social media users' stances towards controversial topics. We analyze users' online interactions that are predictive to their stances and identify the bots within these interactions.

6.1 Introduction

Social media platforms are infested with social bots¹ (automated accounts) that mimic human behavior and can be used to spread inflammatory content with the aim of promoting a specific view or stance (Shao et al., 2018; Bessi and Ferrara, 2016).

Due to the prevalence of bots on social media, humans are not the only players on these platforms, and bots have commonly been used to manipulate views by posting content and interacting with real users (Bessi and Ferrara, 2016; Boichak et al., 2018). For example, these programs were used during the 2016 US presidential campaign to manipulate discussions by spreading content related to the US elections (Rizoiu et al., 2018). In addition, in a recent study by Dunn et al. (Dunn et al., 2020), it shows that bots were used to spread fake news about Coronavirus (COVID19) in social media. All of these factors highlight the need to identify the role bots play in affecting the stance of social media users.

There is no concrete method to analyze the role of bots in affecting social media users' stances (Garimella and West, 2019; Pulido et al., 2018). Nevertheless, there have been several attempts to gauge the effect of bots on various events such as elections (Boichak et al., 2018; Santia, Mujib, and Williams, 2019; Shao et al., 2018). The focus of these studies was to evaluate content diffusion on social networks as a way to measure the influence of bots on public stance towards a topic. Most of these studies evaluated the spread of the misleading and false information by social bots to measure the effect of these accounts on the discussion of an event (Santia, Mujib, and Williams, 2019; Shao et al., 2018); for example, a study by Santia, Mujib, and Williams (Santia, Mujib, and Williams, 2019) evaluated the spread of misleading content on Facebook by bots. Similarly, the work of Shao et al. (Shao et al., 2018) found that bots amplified the spread of fake news within a 10-month period between 2016 and 2017. Previous studies used the spread of bots on social networks as indicators of their effect on social media user's stances. While this method showed that bots are heavily present in social networks, there are still limitations when it comes to identifying whether the presence of bots is correlated with users' stances towards specific topics.

In this chapter, we seek to understand the interplay between bots and support/against stances with respect to a given topic. We study bots' role and define their connection with stance interactions as the signals in the online social network that can be predic-

¹"bots" will be used henceforth to refer to "social bots"

tive for stance towards a given topic, Chapter 4. Our main hypothesis is that if bots exist among the most influential features for predicting the user's stance, then it can be inferred that these bots have a role in pushing and/or reinforcing that stance.

Previous studies on the role and effect of bots on online social networks (OSN) highlight the need to address human interactions to effectively differentiate between automated and real accounts (Abokhodair, Yoo, and McDonald, 2015).

In this chapter, we investigate the following research questions:

- RQ4.1 Do social bots have a presence among the most influential network signals that can predict a user's stance?
- RQ4.2 Is the interaction of social bots with users' stances similar to both those with supporting stances and those with opposing stances? Or do they usually have a more noticeable relationship in a particular direction? Does this change according to the topic?
- RQ4.3 How does the relationship between the presence of bots and users' stance change based on the type of interactions between the bots and users? Do users directly interact with bots by retweeting/replying, or only by being exposed to their content by following bot accounts?

To answer our previous research questions, we performed a large-scale analysis of bots on Twitter. Many studies have indicated a substantial presence of bots on Twitter (Abu-El-Rub and Mueen, 2019; Stella, Ferrara, and De Domenico, 2018), which makes this platform suitable for our study. Then, we built a stance-detection model by using users' interactions as the main features to infer those whose stances were in favor of and against a given topic. We use the stance detection model proposed in Chapter 4, where we incorporate two types of network interactions, direct interactions (IN) and Indirect Exposure (EXP). The (EXP) interactions contain accounts collected from the user's friends list (connection network (CN)), while the direct interactions (IN) include a set of retweets, mentions, and replies to users' tweets.

We applied our experiments on two-stance datasets that contain more than 4,000 Twitter users who had expressed polarized stances towards seven different topics in multiple domains, including the political, social, and religious spheres. The first dataset is the previously used SemEval stance dataset that has been introduced in Chapter 3. The other dataset contains two topics that are related to an emerging events, "Events dataset", which contains tweets related to two events: Brexit and Immigration. We

further explain this dataset in the next Section 6.3. We analyzed those users' networks of interactions and friendships with more than 19 million accounts, among which we identified the bot accounts and the ways in which the users of a specific stance (favor/against) interacted with these bots accounts.

Our findings showed that a relationship between social bot accounts and users' stances does exist, but it is minimal when compared to connections with human accounts, which were more significantly tied to user stance. We also found that the relationship between bots and user stance occurs when users follow the bot accounts and are exposed to their content; this effect was more apparent than the online signals coming from those users who directly interacted with those accounts through retweeting or replying. These findings could help to revisit existing problems in social network analysis, such as understanding the role of social bots in the stance of social media users.

6.2 Related Work

This section provides a discussion of previous work in inspecting bot's effect on social media. Initially, we give some background on Twitter as a social media platform and its policy towards automated accounts, such as bots. Then, we show recent work on measuring the role of bots on the spread of discussions on social media. Finally, we discuss work related to inferring the online signals on social media that are predictive of the stance towards a topic, which is our methodology's primary instrument.

6.2.1 Twitter policy on bots

Twitter is one of the largest online social networks (OSNs). Users can easily create an account, which is public by default, then they can follow any other public accounts without their consent. Only protected accounts, which are accounts that have their posts (tweets) seen only by their followers, are the ones that need explicit approval to follow them.

Unlike many of the social media platforms, Twitter allows accounts to post tweets automatically. This motivated many users and/or institutions to create bots, which are accounts that generate its content automatically and interacts on Twitter based on predefined rules (Seering et al., 2018). Many bots accounts are created for useful causes, such as the Wikipedia edits bots "@EarthquakeBot", which provides updates

about earthquakes that measure 5.0 or more on the Richter Scale, as they happen ². Twitter has a clear policy about the automation of accounts to regulate bots' adoption on its platform ³. One of these rules is to prevent automated accounts from spamming the users or sending unsolicited messages.

Unfortunately, not all automated Twitter accounts (bots) got created for a noble cause. As will be discussed in the next section, some bots get created to spread fake news (Shao et al., 2018) or to create campaigns against election candidates (Bessi and Ferrara, 2016), or to amplify specific stance on a topic (Stella, Ferrara, and De Domenico, 2018). Thus, it became a crucial task for many researchers to build methods to identify bots and measure their spread in social networks. While there are quite many studies in these directions, there is still a limited amount of work to gauge if their role on stance is of any effect. In this study, we fill in the gap by investigating the bots interplay with stance.

6.2.2 Bots' role in social networks

Most of the previous studies assessed the effect of bots by analyzing the spread of these accounts on social media as related to specific events (Abokhodair, Yoo, and McDonald, 2015; Bastos and Mercea, 2019; Ferrara, 2017). For example, a study conducted by RizoIU et al. (RizoIU et al., 2018) used retweet diffusion to analyze the presence of bots in the first US presidential debate in 2016. They used synthetic data and generated an artificial social group of 1,000 users to model cascades of retweets diffusion and to calculate users' importance. The work of Hegelich and Janetzko (Hegelich and Janetzko, 2016) investigated bot activity in the Ukrainian–Russian conflict and concluded that autonomous bot behavior helped spread content. A study analyzed the spread of bots in discussions related to the Syrian civil war by using 3,000 tweets related to the topic (Abokhodair, Yoo, and McDonald, 2015). They found that the growth and content of botnets did not aligned with the bots main behaviour as these bots were spamming the hashtags with topics not related to war. Another study by Bastos and Mercea (Bastos and Mercea, 2019) analyzed the bots behavior in Brexit discourse on Twitter (Bastos and Mercea, 2019). In their study, they used retweets to inspect user-to-bot and bot-to-bot cascade composition. They found that a botnet spread content supporting the "Leave" campaign.

²More examples of interesting and creative online bots are available on Botwiki: <https://botwiki.org/>

³<https://help.twitter.com/en/rules-and-policies/twitter-automation>

The study of Stella, Ferrara, and De Domenico (Stella, Ferrara, and De Domenico, 2018) evaluated the role of bots in spreading negative content according to social media data. In their study, they collected data related to the 2017 Catalan referendum and analyzed the diffusion of negative content by bots. They used Logistic Regression (LR) along with accounts metadata to identify bots accounts. Their results showed that bots increased the exposure to negative content. Along the same line, the study by Luceri et al. (Luceri et al., 2019) estimated stance of bots on social media according to the content they spread.

Another study by Abu-El-Rub and Mueen (Abu-El-Rub and Mueen, 2019) analyzed bot behavior in social media related to the US election and quantified the level of bots and human participation in social campaigns. By analyzing the retweets network, they found that bots' interactions can corrupt social campaigns. Also, Schuchard et al. (Schuchard et al., 2019) examined bots' activities on twitter concerning the US 2016 elections and concluded that bots tend to have a hyper social nature. Along the same lines, Gilani et al. (Gilani et al., 2019) provided a comparison between bots and human behavior with a focus on network activity. They used manual annotations to label the accounts as a bot or not. In their study, they showed that humans have a higher follower rate compared to bots.

Another line of studies analyzed bots behavior on different kinds of platforms, such as Twitch and Wikipedia. For example, Seering et al. (Seering et al., 2018) analyzed the social actions of bot services on the Twitch platform; in this study, they limited their analysis to the service bots provided for Twitch users. Another study analyzed the role of bots on Wikipedia and studied the editing behavior thereof and the effect on human editors (Zheng et al., 2019); they found that the overall human interaction with bots is more in comparison with bots, compared to human—human interactions.

Most of the previous studies examined the effects of social bots by measuring their presence and the spread of their content on social networks. However, there is a gap in the literature to understand if the spread of these bots has a presence within the signals that predict users' stances. Our study extended the efforts to assess the relationship between bots and users' stance on social media by assessing the interplay between bots and users' stances. Moreover, in contrast to previous studies, we provide a fine granularity analysis of bots on polarized stances (i.e., against or in favor). We utilized the advances in stance-detection models using network features to measure bots' presence in the signals that are predictive for stance. Our novel approach states: the more bots that are present among the top predictive features for a specific stance, the stronger the

relation between the presence of bots and a given stance of users on the topic.

6.3 Data collection

To examine the role of bots on users' stances, we utilized datasets that contain ground truth labels for stances toward seven topics. This section provides a description of the datasets used and explains the process of constructing users' networks.

6.3.1 Stance-detection datasets

We used two datasets that contain tweets that are labeled for stances towards seven topics. These datasets are:

SemEval stance dataset. We chose this dataset because it is considered to be one of the most well-known stance dataset that covers topics from different domains. As demonstrated in Chapter 3 section 3.4.1 this dataset contains tweets related to five topics: Hillary Clinton (HC), Climate Change is a real concern (CC), the Feminist Movement (FM), Legalization of Abortion (LA), and Atheism (A).

Events dataset. We created an additional dataset that covered two recent topics: Brexit (B) and Immigration (I). Section 3.4.2 provides further detail of the CD dataset which includes these two topics. These topics were selected because they were one of the viral events at the time we were collecting data . The tweets in this dataset were all selected to be replies to other tweets to have a higher chance of showing a polarized stance as being part of a discussion. We collected 597 tweets on Brexit (B) in February 2019 using the keyword "Brexit." For the topic of Immigration, we collected 1,364 tweets in October 2018 using the following keywords: "immigrant," "refugee," and "border." Tweets of both topics were submitted to the crowd-sourcing platform Appen⁴. We followed the same annotation guidelines used to construct the SemEval stance dataset (Mohammad et al., 2016b), and each tweet was annotated as "favor," "against," by five annotators while taking the majority vote as the final label for each tweet. The inter-annotator agreement between the annotators for Brexit was 73%, and the score for Immigration was 75%; these scores demonstrate a high level of agreement between the annotators, which indicates that different annotators frequently gave the same response (stance) for the same tweet.

⁴previously know as Figure Eight and CrowdFlower. <https://appen.com/>

6.3.2 Collecting users' online networks

For each tweet in our datasets, we collected all network information for its author. For each user, we collected two types of networks, as defined in Chapter 4. The first is $IN_{@}$, which is the interaction network of the user that includes all the accounts the user retweet, mention or reply; and the other is CN_{FR} , which is the connection network of the user that includes the list of accounts the user follows. We used Twitter API to collect users' timelines, which included all the tweets they posted or retweeted in their home-timeline⁵. From the timeline, we extracted all the accounts that the user retweeted, replied, or mentioned to represent the $IN_{@}$. We also collected the friends list (i.e., the accounts the user follows) using Twitter API⁶.

Table 6.1 shows all statistics related to our datasets, including the number of tweets and users for each of the seven topics, which are labeled according to stance, and the number of collected accounts in the interaction and connection networks. As shown in the table, the total number of accounts collected for all the users in our datasets was more than 19 million accounts, which means that on average, each user interacted and/or connected with more than 4,000 accounts in total. The median number of accounts the user interacts with (IN) is 1,288 (average = 2,532), and the median number of accounts the user follows (CN_{FR}) is 602 (average = 2,101).

Our aim in the following was to identify which of those accounts are bots and to understand which of those are shown to have predictive features for specific stances; in this way, we can explore the relationship between bots and user stances online.

6.4 Assessing the role of social bots

This section describes the methodological framework that examined the connection between bots and users' stances in social media. As mentioned earlier, our methodology for measuring the relationship between bots and users' stances is by building a stance classifier using network features and inspecting bots' presence within the most influential features. This section discusses our framework, which includes building an effective stance classifier, extracting the most predictive features for a given stance, and identifying the bots among the accounts.

⁵https://developer.twitter.com/en/docs/tweets/timelines/api-reference/get-statuses-home_timeline

⁶<https://developer.twitter.com/en/docs/accounts-and-users/follow-search-get-users/api-reference/get-friends-list>

Dataset	Topic	tweets (users)	IN@	CN _{FR}
SemEval	Atheism (A)	550 (426)	608,399	740,878
	Climate change (CC)	461 (381)	560,629	524,591
	Hillary Clinton (HC)	670 (511)	1,151,355	1,217,426
	Feminist movement (FM)	524(441)	657,411	371,700
	Legalization of abortion (LA)	670 (490)	978,300	938,184
Events	Brexit (B)	466 (466)	2,129,244	656,864
	Immigrations (I)	1,512 (1,512)	5,567,226	3,274,835
Total		4,853 (4,227)	11,652,564	7,724,478

Table 6.1: The number of tweets per topic in the SemEval and Events datasets with the number of unique users who authored the tweets shown in brackets. The total number of accounts users interacted with (IN@) and followed (CN_{FR}) for each topic.

6.4.1 Stance detection classifier

The first step in our methodology was to build a stance classifier that classifies a given user’s stance as being in favor of or against a given topic.

To create an effective stance classifier, we replicated the current state-of-the-art stance detection model as demonstrated in Chapter 4, which reported the best results—to our knowledge—on the SemEval stance dataset. We used a binary SVM with a linear kernel, and the parameters were tuned using five-fold cross-validation on the training set. In Chapter 4, we showed that a binary classifier that is trained on the two classes of ”in favor” and ”against” while ignoring the ”neither” class achieved a better performance than a three-class classifier. This setup was ideal for our purpose, since we were only focusing on the role of bots on influencing stance and were thus not interested in the ”neither” class ; consequently, we followed this same setup. A stance detection model was trained for each topic separately, which means we trained seven different models for each of our seven topics.

To serve the purpose of our analysis, we focused on network features; namely, the interaction (IN) and connection (CN) network of users, both of which achieved the highest performance on the SemEval dataset, 4.

We trained the stance detection model on two sets of features, and we refer to each of these as follows:

- *IN*, which included all accounts with which each user directly interacts through retweets, replies, or mentions; also includes the website domains the users in-

Topic	A	CC	HC	FM	LA	B	I
IN	71.9	48.2	71.8	61.2	70.3	47.6	55.8
EXP	68.05	48.21	72.98	66.0	66.42	69.2	49.00

Table 6.2: The average F1-score for stance detection on the seven topics in our two datasets.

cluded in their tweets.

- *EXP*, which corresponded to CN_{FR} in Chapter 4, and included the list of the accounts that the user followed. We called it *EXP*, since it represented the accounts the user was *exposed* to by following them, and were thus affected by their content, even without directly interacting with the content thereof by liking, replying, or retweeting.

For each topic in the datasets shown in Table 6.1, the labeled datasets were split into 70% /30%, following the same split reported by the SemEval stance dataset (Mohammad et al., 2016b) for training and testing our classifier. In the SemEval dataset, we used the same split for training and testing that was used in Chapter 4. For the *Events* dataset, we applied a random split to training and testing with the same split percentage—70% and 30%—for training and testing, respectively. The stance classifier was separately trained for each topic twice: once by using the (*IN*) features of each user, and another by using the *EXP* features thereof. For evaluating the classification performance, We used the SemEval stance-detection official evaluation script to calculate the F1-score, Chapter 3, section 3.6.2.

Table 6.2 shows the performance on the seven topics reported in the F1-score using the script provided by the SemEval task (Mohammad et al., 2016b).

6.4.2 Extracting the most influential features on stance

To assess the extent of the relationship between bots and users' stances, we analyze the most effective features for the stance prediction model. For each polarized stance (favor or against), we use the weight of the coefficient generated by the stance model to identify the set of the most influential features on the stance prediction. These features are extracted from the feature set which contains the accounts and domains (URLs) the user interacts within the (*IN*) feature set, and the accounts the user follows

in the (EXP) feature set. We use the top 1,000 most influential accounts for the stance prediction model from each feature set, excluding the domains from the IN features in our analysis, since our focus in this study is on bots' connection to user stance.

In the next section, we inspect the population of bot accounts that exist in those 1,000 predictive accounts for user's stance, and compare their population to other accounts.

6.4.3 Identification of bot accounts

There is a large body of work focused on the development of techniques to detect bots in social media (Davis et al., 2016; Puertas et al., 2019; Santia, Mujib, and Williams, 2019). The work of Puertas et al. (Puertas et al., 2019) used a multilingual classification model to identify bot accounts based on the content of their posts. One of the most popular bot detection APIs is Botometer®⁷ (Davis et al., 2016; Yang et al., 2019), which provides a robust method to detect the existence of bots in social media. Botometer® uses a Random Forest classification algorithm to classify tweets as bots based on 1,000 features that were extracted from users' meta data along with tweets timeline. The classification score ranges from 0 to 1, where 0 indicates the likelihood that the account is human and 1 indicates that the account is likely to be non-human ("bot").

The Botometer® API has been used in various studies to detect the existence of bots in the network (Rizoiu et al., 2018; Varol et al., 2017; Broniatowski et al., 2018). In the study conducted by Rizoiu et al. (Rizoiu et al., 2018), the Botometer® API was used to analyze the role and influence of bots on social media in the 2016 US Presidential Debate. Another study by Broniatowski et al. (Broniatowski et al., 2018) estimated the bot scores of Twitter accounts that spread content related to the vaccine debate on social media. Along these lines, we used Botometer® in our study to detect the bots in the users' networks in our dataset.

To identify the bots in the set of predictive accounts extracted from the stance detection model, we used the Botometer® API (Davis et al., 2016). This API generates a score $\in [0 - 1]$, where 0 indicates the account of a real user, and 1 suggests the strong likelihood of a bot.

Sometimes the Botometer® API generates an error message. This happened when it failed to access the tweets of an account because it is deleted, suspended, or pro-

⁷<https://botometer.iuni.iu.edu/>

tected. We considered protected accounts to be human accounts, since it was unlikely that a bot would restrict its tweets to only its followers (Rizoïu et al., 2018). While suspended accounts could be suspended because they were bots, Twitter can also suspend an account if the user of that account violates the platform rules⁸. Common reasons to suspend a Twitter account includes abusive tweets, spamming, or if the account has been hacked or compromised. For the previous reasons, we did not consider the suspended accounts in our dataset to be bots; instead, we treated these accounts as "unknown" and labeled them as "deleted." Therefore, the deleted and suspended accounts in our dataset were label as "deleted".

For accounts that have a low botometer score, which are most likely to be non-bots (i.e., human accounts), we wanted to make a distinction between famous and normal accounts, since it might be expected that influential accounts will have a more prominent relationship with users' stances than normal accounts. According to the research conducted by Cossu, Labatut, and Dugué (Cossu, Labatut, and Dugué, 2016), the authors postulated that an account was more influential when it had more followers. Thus, we further classified the non-bot accounts according to the number of followers thereof into three categories: ultra-famous, famous, and normal. According to Twitter users statistics⁹, only 0.05% of Twitter accounts have more than 10,000 followers; thus we label them as ultra-famous; the famous accounts are those with a number of followers ranging between 1,000 and 10,000, and which applies to 2% of Twitter users; and finally, the normal accounts were those with fewer than 1,000 followers, which applies to 98% of Twitter users.

6.5 Results and Analysis

In this section, we assess the role of the social bots in detecting online stance by analyzing the top influential accounts that were the most predictive toward polarized stances regarding each topic.

6.5.1 The distribution of bot scores of the most influential accounts

Figure 6.1 shows the distribution of Botometer scores for the top 1000 accounts that are most predictive for stance in our dataset, for both the interaction and exposure

⁸<https://help.twitter.com/en/managing-your-account/suspended-twitter-accounts>

⁹<https://sysomos.com/inside-twitter/twitter-statistics/>

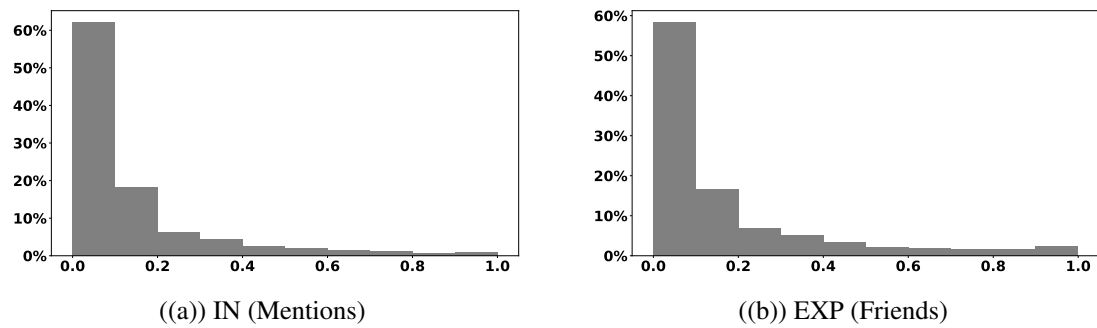


Figure 6.1: Botometer score distribution of the top 1000 accounts that are predictive to stance for both networks.

network features (IN and EXP)¹⁰. The scores generated by Botometer are within a range $\in [0 - 1]$, where 0 indicates the account of a real user, and 1 suggests the strong likelihood of being bot.

As shown in Figure 6.1, most of the accounts that are predictive to stance have a low bot score, where the majority of accounts have scores between $[0 - 0.2]$, indicating that these accounts are most like to be real people. Only very few accounts have high bot scores (≥ 0.6). This indicates that most of the accounts that have a role in predicting users' stance are for real people.

To enable a more in-depth analysis of accounts that are more likely to be bots, we focused on those accounts that got a score of over 0.6, which indicates high likelihood of being a bot, which is the same score used in previous studies to analyse bot behaviour. (RizoIU et al., 2018; Ferrara, 2020). In these studies, an account was classified as a bot when the Botometer® score exceeded a threshold of 0.6, where they showed that this score decreased misclassification and improved the overall bot-detection accuracy (RizoIU et al., 2018). We followed the same setup and used the same threshold¹¹.

The next section provides a further analysis of these accounts that are likely to be bots given their high bot score and compares them to human accounts (accounts with low Botometer score) and deleted accounts.

¹⁰Appendix C shows the Botometer scores distribution on topic level

¹¹We also examined a threshold of 0.5, but found that it increased misclassification of the human accounts as bots without improving the detection of new bots.

6.5.2 The role of social bots on stances

For each type of feature (i.e., IN and EXP), we show the percentage of likely-bot accounts alongside other types of accounts for each topic.

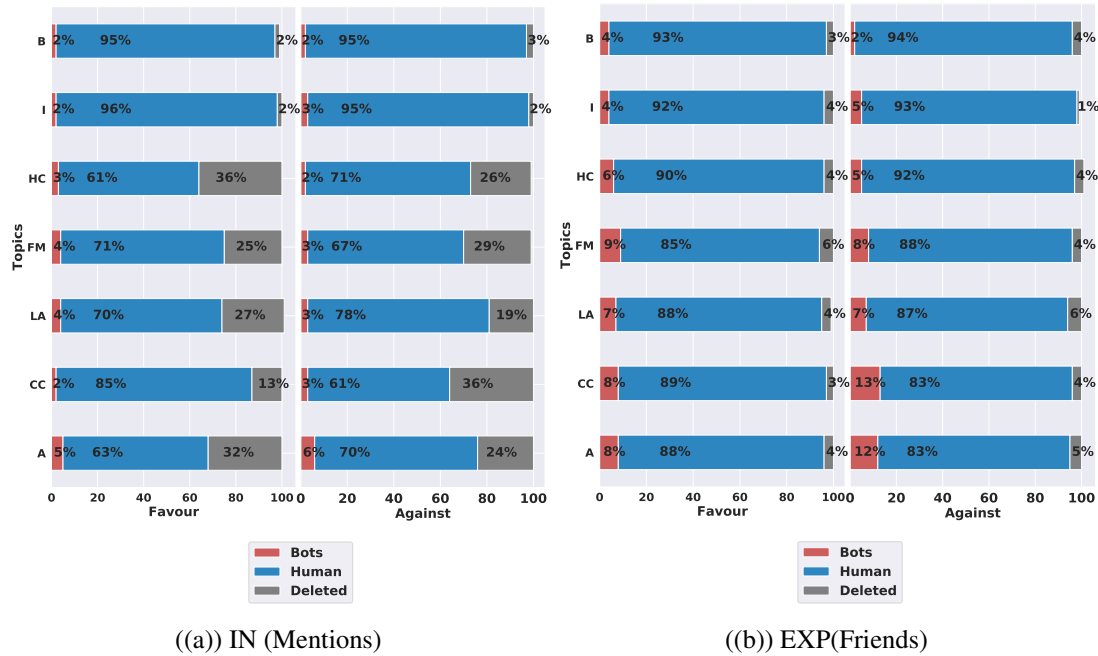


Figure 6.2: Distribution of social bots for each topic in the top 1,000 most predictive accounts for polarized stances using direct interaction (IN) and indirect exposure (EXP) features.

Direct interactions (IN). Figure 6.2 shows the percentage of social bots on the in-favor or against stances with respect to the top 1,000 IN features for each topic. The results show the prevalence of real accounts in the set of most influential features on the stances, compared to the minuscule existence of socialbot accounts. This trend is consistent in each topic with respect to both in-favor and against stances. The bots had an existence not exceeding 6% of the overall set of influential accounts in each topic, while non-bot accounts constituted the majority of the most influential accounts, reaching higher than 95% for some topics. As shown in figure 6.2(a), some of the accounts in the top 1,000 were deleted by the time we inspected them, especially those related to the SemEval stance dataset topics, since the data was more than four years old. This is one of the limitations of working with Twitter, since we cannot retrieve information from those accounts after deletion. This kind of limitation is well known in the online social network studies (Boichak et al., 2018; Ferrara, 2017). Nevertheless, we still had tweets wherein these accounts were mentioned in the collected users'

timelines, which allows us to provide further analysis to these accounts.

Indirect exposure (EXP). Figure 6.2(b) illustrates the percentage of bots in the favor or against stance with respect to the top 1,000 EXP features for each topic. Again, the percentage of bots is minimal compared to human accounts. However, it is worth noting that bots constitute more population in the EXP network compared to IN, where it reaches 12% and 13% in some cases (Climate change and atheism topics). This suggests that being exposed to bots' posts might be more strongly tied to users' stance than direct interactions with bots.

Furthermore, these bots that people follow have a stronger connection to the *against* stance of some of the topics compared to the *favor* stance. For instance, users with against stance towards Atheism tend to be affected by bots accounts more than users supporting Atheism (12% against vs. 8% in favor).

The same trend can be seen in Climate change and Immigration. This demonstrates that bots can have a larger relationship to one stance direction compared to the other.

6.5.3 Magnitude of the bots' role

The previous analysis showed that the majority of the top 1,000 predictive accounts of predicting stance for all the topics was for humans. However, a small proportion of bots might still be the most influential among those 1,000. Thus, in this section, we provide a more in-depth analysis to the distribution of bots in the top N predictive accounts, where N ranged between 10 and 1,000. In addition, we analyze the type of human accounts according to how famous they are.

Figure 6.3 illustrates the average distribution of different types of accounts on each polarized stance extracted from the IN and EXP networks averaged across all topics. We noticed that the distribution of bot accounts constituted the lowest percentage across all values of top N features. In fact, the average distribution of bots never exceeded 10% at any point. This was consistent across both networks and for both stances. The ultra-famous accounts, which were the human accounts with more than 10,000 followers, constituted the majority of accounts; they consistently constituted more than 50% across all values of N over all stances and networks, and their influence reached over 70% in the top 10 features for the EXP against stance. This means that following these accounts related to users' stances being against a given topic.

Further detailed results for each of our seven topics are presented in Appendix C, where we show variations across some of the topics. For example, for Brexit, the

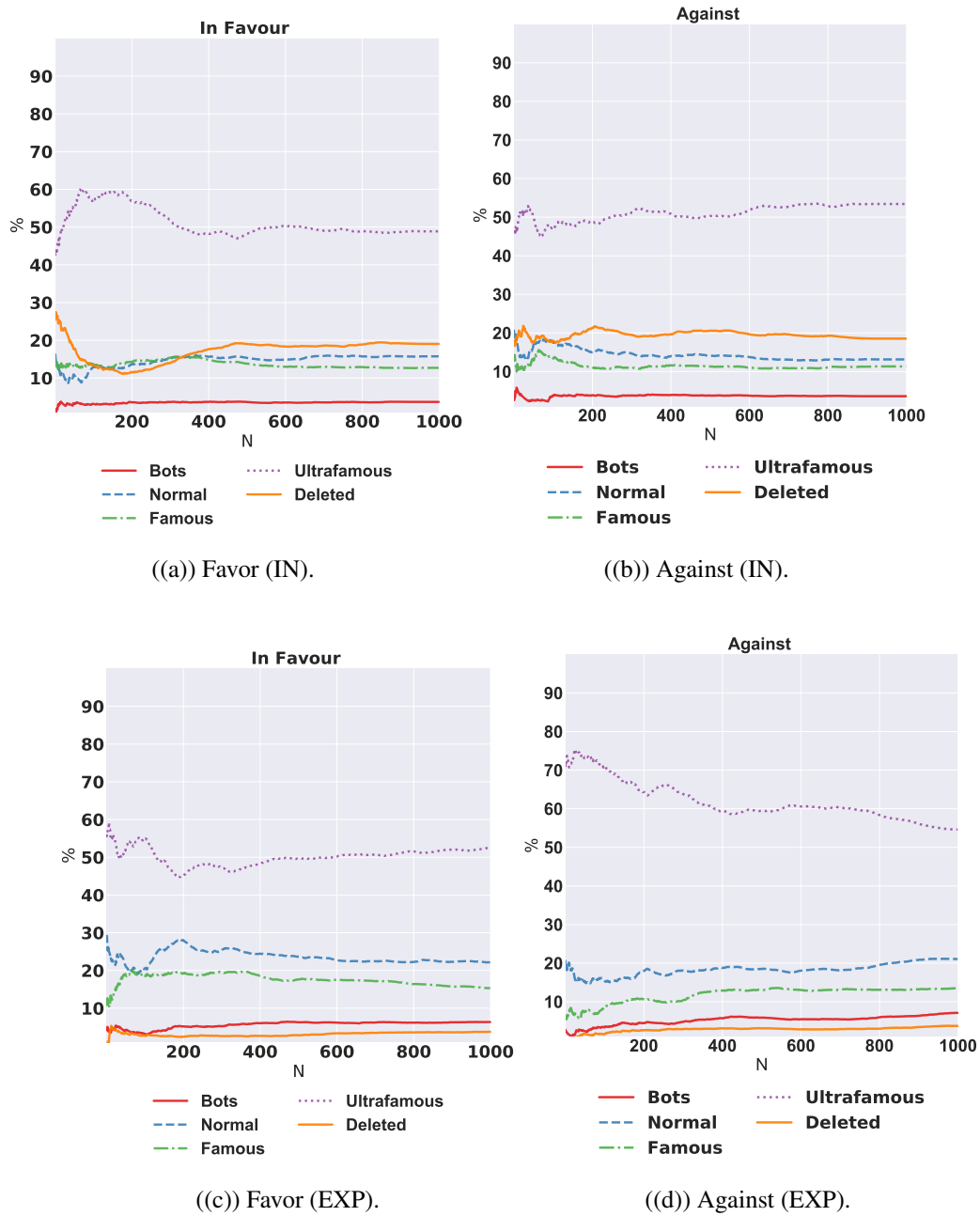


Figure 6.3: The percentage of each account type (X-axis) in the top N (Y-axis) influential accounts in predicting the Against/Favor stances in direct interactions (IN) and indirect interactions (EXP).

ultra-famous users had a noticeable connection to the in-favor stance, reaching approximately 75% of the top 100 accounts, while the ultra-famous users only constituted 25% in the against stance. Moreover, for the against stance in EXP interactions, these accounts showed a sizable presence in the top 100 features for six topics: atheism, cli-

mate change, the feminist movement, Hillary Clinton, immigration, and Brexit. For the legalization of abortion, the normal accounts with fewer than 1,000 followers had the most presence in the top 800 features of people who were opposed to the legalization of abortion.

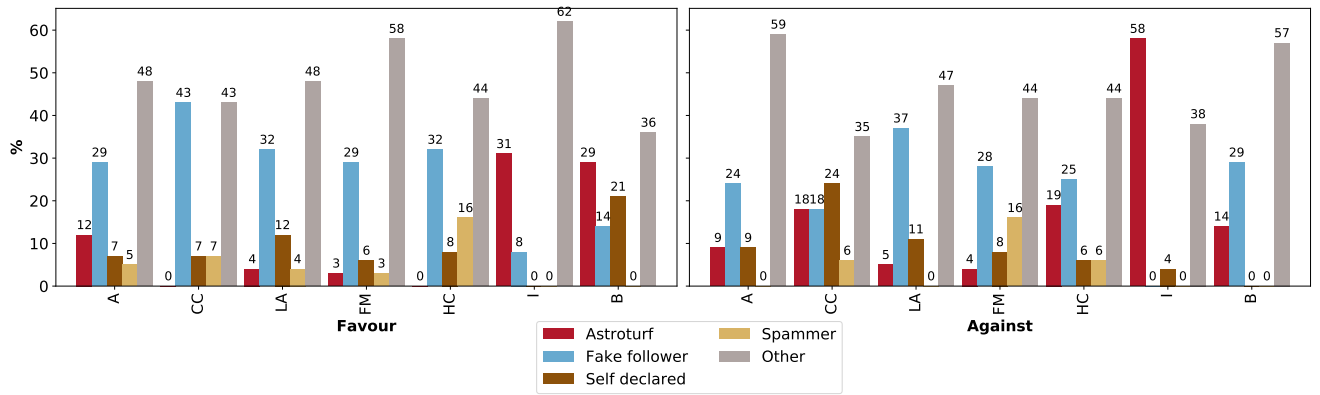
Our analysis shows that bots have some role in relationship to online stance of being in-favor/against a topic, however to a much smaller degree than what we expected in our first hypothesis (H1). The relationship is minimal compared to that with human accounts, especially the ultra-famous accounts, which had the most association with users' stances by far. We also found that bot accounts that people follow and are exposed to their content (EXP) has more influence (presence in the top features), than bots with which users directly interacted. We applied a statistical significant test using Pearson's chi-squared test between the distribution of bot accounts in the IN and EXP and found that bots presence in the EXP network is statically significantly higher than IN for all stances in most of the topics with $p - value < 0.001$, except the Brexit and immigration topics, where both had the least number of bots (only 2–3%). Table C.1 in Appendix C shows the full values of the chi-squared test per topic and stance. This result confirms our third hypothesis (H3) that people stances can be affected indirectly just by getting exposed to bots content, as we actually showed that EXP network affects users' stance more than IN network.

6.5.4 Properties of the influential bot accounts

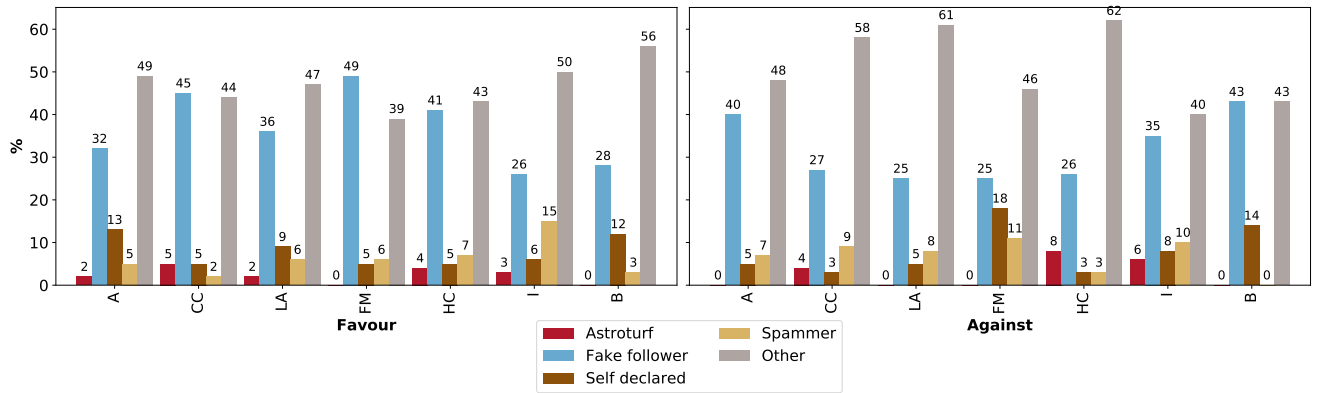
We further analyze the properties of the bots identified within the most influential accounts on predicting users' stance, where we check their types and the number of followers they have in comparison with the human accounts.

Types of the influential bots. The new version of Botometer API (V4) provides the type of bot based on of six categories¹². These categories are: *Astroturf*, *Fake follower*, *Financial*, *Self declared*, *Spammer* and *Other* (miscellaneous). The *Astroturf* bots are political bots and accounts involved in follow trains that systematically delete content. The *Fake follower* bots are bots purchased to increase follower counts. *Financial bots* are the automated accounts that post using cashtags. The *Self declared* are bots labeled using botwiki.org, which is a website that keep track of useful and creative bot accounts that self-declare themselves as bots. While the *Spammer* are automated accounts labeled as spam bots from several datasets. Bots labeled as '*Other*' are the

¹²<https://cnets.indiana.edu/blog/2020/09/01/botometer-v4/>



((a)) IN (Mentions)



((b)) EXP (Friends)

Figure 6.4: Distribution of social bots types for each topic in the top 1,000 most predictive accounts for polarised stances using direct interaction (IN) and indirect interaction (EXP).

miscellaneous other bots obtained from manual annotation or reported by other users.

We used Botometer V4 to analyse the types of bots we identified in the influential accounts. Figure 6.4 shows the distribution of bots types (IN) and (EXP) networks. It can be noticed that most dominate bots are the bots of type *Others*, that are obtained from manual annotation and user feedback. The *Astroturf* bots constitutes most of bots in the direct interactions networks that influence the against stance towards immigration and favor stance towards Brexit. Nonetheless, *Astroturf* bots shown to have the minimal presence in the indirect interactions (EXP). Bots that are identified as *Fake followers* have good presence in most topics, especially the exposure network. Overall, *spammer* bots constitute the minimal percentages over direct and indirect interactions, while *Financial* bots have no presence in the most influential accounts. These results show that bots that have an interplay role with stance are mostly the ones that get re-

IN				
Account type	Favour		Against	
	bots	humans	bots	humans
Normal	57.73%	16.52%	51.57%	12.53%
Famous	19.58%	16.70%	14.73%	12.55%
Ultra-Famous	22.68%	66.77%	33.68%	74.90%

EXP				
Account type	Favour		Against	
	bots	humans	bots	humans
Normal	63.39%	21.71%	59.25%	20.63%
Famous	22.00%	16.66%	20.07%	14.68%
Ultra-Famous	14.59%	61.61%	20.66%	64.68%

Table 6.3: Distribution of bots and human based on followers. The Ultra-famous accounts $> 10,000$ followers; The famous accounts are those with number of followers ranging between 10,000 and 1,000; The normal accounts $< 1,000$ followers.

ported by normal users, while political bots (astroturf) still have some role, especially in stances on political topics.

Followers of influential bots. As our analysis shown, ultra-famous accounts are the most influential in predicting stances. Thus, we further analyzed the bots' number of followers to understand to compare them to the influential human accounts. Table 6.3 shows the split of the identified bot accounts by their number of followers as normal, famous, and ultra-famous. As shown, the majority of bots (50-60%) have less than 1,000 followers. However, around 30-40% of them still have large number of followers, especially for those bots influential on the against stance in the IN, around 34% of them are considered ultra-famous accounts by having over 10,000 followers. This shows that some of the bots that are popular on Twitter.

6.5.5 The context of the influential bots

In this section, we analyze the context of interactions between bots and users for each stance. We explore some examples of where influential bots appeared in users' timelines to understand the possible link between the bots and users' stances. We also checked the type of some of the bots that the users followed to estimate the exposure that might have affected their stance.

#	T	Stance	Example tweet
1	A	Against	RT @FollowDMs: Follow everyone who retweets this
2	A	Against	RT @JesusNarrowWay: 1 Peter 4:18, If it is hard for the righteous to be saved, what will become of the ungodly and the sinner?
3	HC	Against	RT @VoteHillary2016: Donald, are you talking about the 70K votes we lost in 3 states or the nearly 3 million popular votes you lost despite
4	FM	Against	'So @ForgetFeminism according to this...99.99% of the feminists I talk to are NOT "feminists".ll let them know.#WomenAgainstFeminism'
5	A	Favor	RT @BibleWisdoms: There's one Lord, one faith, one baptism, and one God and Father of all - Ephesians 4:5-6
6	HC	Favor	RightOn! @Timoniumbill: @ReadyForHillary Mrs. Clean. http://t.co/xBh7FrjZXh #OhHillNo #WakeUpAmerica #StopHillary2016
7	I	Against	RT @cookequipman1: AMERICA'S VET TRAIN #ConnectingAmericanVets #MAGAveteran

Table 6.4: Sample of tweets and the context of social bot interactions in relation to stance and topic.

The direct interaction with bots. Table 6.4 presents a sample of the tweets generated by the bots that were the most predictive in the IN features. We found these bots in the users' timelines as retweets, replies, or mentions of the tweets. In general, bot interactions with social media users have three forms: 1) bots with content that aligned with the user's stance; 2) bots with content that disagreed with the user's stance; and 3) bots with content that had no relation to the user's stance. For instance, the bot account "@FollowDMs" was one of the influential accounts in predicting the against stance to atheism, yet this account had no relation to the topics of atheism or religion (see Example 1). Additionally, regarding bot accounts that had no direct relation to immigration, the "@cookequipmant" account was one of the most influential accounts for predicting the stances related to the topic of immigration (see Example 7). Furthermore, Example 5 shows that users who supported atheism tended to directly interact with bot accounts that contradicted their stance. For example, one of the influential bot accounts in predicting the in-favor stance towards atheism was the one that promoted religious content, "@BibleWisdoms." Additionally, Example 3 shows that the account "@VoteHillary2016" was a bot account that supported Hillary Clinton, yet it was one of the accounts that had a strong effect on predicting the against stance to Hillary Clinton. Moreover, users with an against stance to atheism interacted with religious bot accounts that promoted religious content, such as "@JesusNarrowWay" (see Example 2).

The indirect exposure to bots. We also analyzed bots accounts that were the most predictive in the EXP features, which were the bots that the users in our datasets were following. Table 6.5 presents the top three bots that influenced the against and in-favor stances for each topic. In general, social media users tended to follow bots that aligned with their leanings. For example, people with an against stance toward atheism tended to follow automated accounts with religious content, such as "2ayaat" and "RTALD3OAH." The same observation was made for users who supported Hillary,

T	Favor	Against
A	HaginQuotes, RCSproul, warpawsiraq	2ayaat , lilxstyles, RTAL_D3OAH
CC	Smartassy4ever, jtd_gameon12, bigboater88	AllAmericanGirI, SassyCon, Moonbattery1
HC	WhatHillaryAte, bluenationuntD, stylebysassys	saynotogop, humoryoulike, UniteBlueSC
FM	geekfeminism, onlyminionquote, tomily4	stopbrutality, FeministShit, SC2TopReplays
LA	sucessfultips1, JohnGaltTCMC, SMNW_YRC	TheKeyisPrayer, prolife321, myjesus123
B	UKPollingLive, watching_eu, Brexit_WestMids	moggality, mosthauntedlive, britainsmilhist
I	RealBarcaBooks, RomanCatholic36, umustknowthis1	milagrovargas14, fridayfeeing,mrmarkel

Table 6.5: Top bot accounts in indirect interactions for each stance towards the seven topics.

who tended to follow accounts that confirmed their leanings, such as *"WhatHillaryAte"*; this account was an automated account with retweets and tweets that amplified support to Hillary. Additionally, users with stances that supported the feminist movement followed accounts that promoted the feminist movement, such as *"geekfeminism."* As it relates to Brexit supporters, accounts such as *"Brexit_WestMids,"* which promoted Brexit through tweets and retweets, which was one of the most effective accounts in the friends list to predict the in-favor stance towards Brexit. It is worth noting that, in general, the most effective bots for stance prediction had no direct relation to the topic related to the stance. This can be seen in the top three bots that interacted with users who held an against stance toward climate change; these accounts tended to cover a variety of political subjects that had no relation to climate change. For example, the account *"AllAmericanGirI"* posted news tweets that were related to the Conservative party. Furthermore, users with an against stance toward Brexit tended to follow bots that distributed content about political news, but had no direct relation to the withdrawal of the United Kingdom from the European Union.

6.6 Inspecting the deleted accounts

On average, the deleted accounts constituted approximately 19% of the overall influential accounts in direct interactions and about 11% of the indirect interactions. One of the limitations in our previous analysis was failing to analyze these deleted accounts. For some of the topics, the number of the deleted accounts in the top 1,000 was over 30%; we cannot confirm whether these accounts were bots or real users. This limitation is usually found in the studies of social bot behaviors as a result of collecting tweets in the aftermath of an event (RizoIU et al., 2018; Luceri et al., 2019; RizoIU

et al., 2018; Shao et al., 2018; Howard and Kollanyi, 2016). These deleted accounts have presented a hurdle in many bot-detection studies (Rizoiu et al., 2018; Luceri et al., 2019); as these accounts no longer exist on the Twittersphere, so it was difficult to retrieve the needed information for these accounts to examine the bot behavior of the account. For example, in a study conducted by (Luceri et al., 2019), the dataset was composed of approximately 99% suspended accounts. In our work, since we focused on user stances and using the set of influential accounts, the percentage became much lower, compared to previous studies of bots on social networks. The deleted accounts in friends networks (EXP) constituted a much lower percentage in comparison with direct interactions (IN). This is due to the fact that the accounts collected from the direct interactions were extracted from each user's timeline, which may have contained obsolete mentions, while the friends set tended to only contain the accounts that existed at the time of the collection. In an attempt to overcome part of this limitation, at least as it related to the deleted accounts in the IN features, which had the highest deletion percentage, we decided to manually inspect all the tweets where they were mentioned in the collected users' timelines; then, based on this tweets, we decided whether they were bots or not. Since this process was time consuming, we considered all the deleted accounts in the top 100 of the influential features of the IN features. As Figure 6.3 shows, these trends can be spotted within the first 100 features of the direct interactions. We used the same annotation guideline of the Varol-2017 (Varol et al., 2017) to label the deleted accounts as bots or not. The annotation guideline of the bot-detection study by (Varol et al., 2017) was based on inspecting the account's profile page and looking for common flags, such as a stock profile image or retweeting that occurs within seconds. As these deleted accounts had no Twitter profile information, inspecting these accounts' profile pages was not applicable in our case. Since there was no unified rule to label an account as a bot, we further retrieved a set of tweets from stance dataset where the users interacted with these deleted accounts. These tweets provided additional clues to label the set of deleted accounts and inspect their behavior.

After manual inspection, we found that some of those accounts were likely bots. Table 6.6 presents the amount of existing bots and deleted bots in the top 100 IN features. In general, bots constituted less than 11% of the top 100 features of in-favor and against stances. The topic of immigration had the most proportion of bots that interacted with the in-favor stance. As it related to against stances, the legalization of abortion and atheism contained the largest amounts of bots, compared to other topics.

T	FAVOR				AGAINST			
	deleted	deleted bots	existing bots	total bots	deleted	deleted bots	existing bots	total
A	8	2	3	5	20	8	0	8
CC	8	2	2	4	15	0	0	0
HC	33	8	1	9	20	2	2	4
FM	23	1	1	2	27	2	5	7
LA	25	4	5	9	10	3	5	8
B	12	5	2	7	9	2	5	7
I	8	3	7	10	7	3	4	7

Table 6.6: The number of deleted accounts and the expected bots in the top 100 influential accounts on stance prediction.

Table 6.7 shows a sample of tweets that demonstrated the characteristics of the deleted accounts. Some of the deleted accounts had the term "bot" as part of the user name, such as in Example 3. In other cases, the account name indicated the behavior of the account, such as "@theism_sucks" (see Examples 4 and 5). User interactions with this account were conducted in the sense of mentioning it to defend their religious perspectives. Some accounts had limited content in our dataset such as in Example 6. In this example, the account "@BarbietheBrain" appeared in a retweet with other accounts as a means of promoting these accounts. We considered such situations as promoting an account by spreading automated content and labeled the account as a bot¹³. Other deleted accounts had tweets that showed somewhat personal messages such as in Example 7. The account "@coolredmac" was a suspended account, as this account had hateful tweets, such as those in Example 8. In this case, we labeled this account as not being a bot account. Other accounts had normal content based on the retweet behavior in our dataset, such as in Example 9. The account "@PETTYMAMII" was a suspended account which has non-hateful tweets when we retrieved the account's timeline tweets from our dataset.

One of the obvious indicators of bot behavior was the vast amount of retweets that showed the content of the account, such as those in Examples 10, 11, and 12. These examples showed retweets to a bot account "@LiveActionFilms" interacting with the against stance toward the legalization of abortion. This account was suspended because of spreading negative messages.

¹³In Example 6, a combined account with "@BarbietheBrain" posed bot behavior based on their profile characteristics

	T	S	Type	Tweet
1	CC	-	NotBot	@_PinealGland: 1984 https://pic.twitter.com/DgmnISvYON " one of the best books ever
2	HC	-	Bot	@srtalbot2 http://t.co/fjr9IeSKak
3	A	-	Bot	RT @ArchbishopYoung: "Today is your day, your mountain is waiting, so get on your way." - Dr Seuss #Quote
4	A	+	Bot	@X @theism_sucks we christians dont want dark matters to rule our world. we love the #Light #happiness #God'
5	A	+	Bot	@2ManyOfUs @theism_sucks pettiness? #bible speaks the truth. #owned again #atheist sucks LOL'
6	B	+	Bot	RT @Silentwoo: @IrishVol69th @alley167 @BarbiethBrain @LisaNiebs @NeensCa@heyitsCarolyn @Sekusa1 @ON11...
7	B	+	NotBot	@X @coolredmac Well said, Sir.
8	B	+	NotBot	RT @coolredmac: Which is why she is no longer prime minister Emmanuel Macron praises Theresa May for being loyal and respectful to EU
9	FM	+	NotBot	RT @PETTYMAMII: Not seeing your best friend for a long X time really hurts .
10	LA	-	Bot	RT @LiveActionFilms: Paul: "The right to life and freedom of religion preexist government." #VVS14 #prolife'
11	LA	-	Bot	RT @LiveActionFilms: "Humanizing," @PBS? What is human, anyway? Watch the video:X #AfterTiller #abortionaccess'
12	LA	-	Bot	RT LiveActionFilms: Our latest video showing PPact dangerous #SexEd for kids was featured on OReilly last night!"

Table 6.7: Sample of tweets that from accounts that interacted with deleted accounts in the top 100 features of (IN). We used "X" to mask some users accounts and hide sensitive content.

6.7 Verifying the bot/non bot accounts

In order to verify the reliability of Botometer in detecting the bots and non-bot accounts, we verified the propriety of the top 10 accounts for each topic/stance and identifying the likely bot accounts. We inspect the type of the accounts and measure the Cohen's kappa score between the Botometer and annotation labels to gauge the reliability between the two labels. We used the same annotation guideline of the Varol-2017 (Varol et al., 2017) to verify the likelihood of bot. Also, we used Bot-Detective API Kouvela, Dimitriadis, and Vakali (2020) to provide further explainable hints for bot-like accounts, that helps us to provide a ground truth labels by using extra information beyond inspecting Twitter page. Table 6.8 provides some examples of accounts and examples of explanations provided by Bot-Detective in identifying the bot-likely accounts. We found a high alignment between manual annotations and Botometer labels in identifying human accounts with Cohen's kappa score equal to 68.81%, which indicates a substantial agreement between the manual annotations and Botometer. Even in cases where the account seems to be a non-personal account, using Bot-Detective helps in verifying those accounts. For instance, the account *@hqtriviafans* is a fan page for the trivia game show. Example 3 in table 6.8, shows that this account has a high likelihood to be human as the average number of characters per tweet (72.75). Bots usually have 143.7 characters on their tweets. Although, for some bot accounts, the score was on edge, even for the annotators. This is due to the fact that some of these accounts are mostly having low tweets with the default setting. In these cases, we use Bot-Detective to provide explanations based on non-profile information and further inspect the type

	T	Type	Verify	Account with explanation
1	A	NotBot	✓	@RichardDawkins This account is verified. Almost always, this means that the account belongs to a non-bot user.
2	I	Bot	✓	@GOTGeekX This account's URL per word ratio for each tweet, is suspiciously high.
3	I	NotBot	✓	@hqtrivalfans Normal average number of characters per tweet (72.75). Bots usually have 143.7 characters on their tweets.
4	HC	Bot	✗	@laura.beene the average liked tweets is normal
5	LA	Bot	✗	@saysmysister This account uses symbols rarely (11.53 symbols per tweet). Bots usually have 21.2 symbols per tweet, on average.
6	LA	Bot	✗	@thomash Normal number of hashtags on tweets. Bots usually have 3.48 hashtags on their tweets and this account has 0.

Table 6.8: Sample of verified accounts with explanation from Bot-Detective tool.

of those accounts. For instance, the account *@GOTGeekX*, has a score of 3.7 out of 5 using Bot-Detective, and 0.96 in Botometer. Using the explanations generated by Bot-Detective, the account is highly likely a bot, considering that the URL per word ratio for each tweet is suspiciously high (table 6.8, example 2).

6.8 Discussion

Given the prevalence of bots in social media, it is crucial to examine the role these accounts play in affecting the online users' stances and to understand the interaction behavior of these accounts. Measuring the factors that relate to people's stances in social media is a complex process that is influenced by various behavioral signals (Lee et al., 2010; Cha et al., 2010) . Motivated by this challenge, we investigated the role of bots using a gold-standard stance-labeled dataset that contained real users' stances on seven topics; this dataset contained events and topics that covered three main domains (i.e., politics, religion, and social aspects). In this study, we extended our understanding of the relationship between bots and the stances of social media users, and we highlighted various implications for bot studies.

6.8.1 Bot and human effect on stances

To answer the first research question and to assess the association between bots and stance, we analyzed the most influential accounts to predict users' stances, and we inspected the presence of bots among these accounts. We showed bot and human distributions were within the top 1,000 most predictive accounts for their stances concerning seven topics. Overall, while bot accounts were present in the top influential accounts in predicting the stances thereof, the bots had the lowest percentage, compared to human accounts, as shown in Figure 6.2. This result while it confirms our first hypothesis (H1) where bots have a presence in the top features, this presence is considered much less than what we expected. This finding places an emphasis on the noticeable connection between human accounts and a given stance, compared to that

of bots. Our results align with the recent study by Dunn et al. (Dunn et al., 2020) who investigate the effect of bots in-comparison with people in social media dataset related to COVID19. They found the role of bots on spreading fake news about the anti-vaccine is limited.

Moreover, we showed the magnitude of the effect of the top 1,000 accounts on predicting the stances related to three kinds of real user accounts (i.e., normal, famous, and ultra-famous), as shown in Figure 6.3. The noticeable link between the *ultra-famous* accounts and stance formation can be observed in the first top 10 accounts that influenced the given stance. This finding does not align with the "million followers fallacy" theory (Avnit, 2009), which was confirmed for Twitter by (Cha et al., 2010). Throughout this study, we showed that the generalization of the followers' theory is not applicable in the realm of measuring the influence and connection to stance.

Furthermore, we provide a finer granularity analysis of the role of the bots on the topic level (see Figures C.4 and C.3 in Appendix C). It can be observed that the relationship between the ultra-famous users and the given stance represents the general trend on the topic level.

6.8.2 The link between bots and supporting versus opposing stances

When addressing the second research question, we noticed that the role of bots on the supporting and opposing stances was relatively different for a majority of the topics, which aligns to our second hypothesis (H2). This can be seen in the proportion of bots that influenced the stances, as shown in Figure 6.2, even though the bots presence in the two topics of atheism and climate change was sizable on the against stance, compared to the in-favor stance. However, by inspecting the bots in the friends set (EXP) for climate change, we found that most of these automated accounts had no direct relation to climate change. As for the other topics, there was at least one bot account in the top three accounts in friends sets that were related to the topic of the stance. This finding indicates that bots can have a greater role on a specific stance type than the others for some topics. For instance, users with stances that opposed atheism tended to follow and interact with bots that had religious content.

The same was observed for the friends accounts (EXP) that influenced the against and in-support stances towards the legalization of abortion, which had approximately 7% bots. Furthermore, there was a noticeable difference in bot distribution at the topic level. The presence of bot accounts was sizable in the direct interaction (IN)

and indirect exposure (EXP) , in atheism, as is shown in Figure 6.2. The fewest bot accounts were seen in Brexit, where bots constituted approximately 2% of the overall interactions. When we further inspected the type of bots that influenced the stance toward atheism, we noticed that these accounts had a religious theme that promoted faith, which supported the specific type of the stance. Similarly, the bots that influenced Brexit stances tended to have a political theme and a focus on news related to the withdrawal of the United Kingdom from the European Union, such as *watching_eu* and *Brexit_WestMids*.

Moreover, we inspected users' behavior when interacting with bots. We showed that users tended to directly interact with bots that had a stance that was different from theirs (see Examples 3 and 5 and Table 6.4). This indicates the simple direct interaction to a bot's content does not have a direct relation to a user's stance. This behavior can be supported by the backfire effect (Nyhan and Reifler, 2010), which means that exposure to this kind of content has a negative effect on people's stances. This contradicts the main goal of creating social bots that aim to spread negative content to manipulate views towards a topic. This finding helps us to gain a better understanding of the effect of social bots on users' stances on social media. It is worth noting that social media users tended to follow bots that aligned with their leanings. This can be seen in the top accounts that influenced the against stance towards atheism, where users tended to follow religious accounts, such as *2ayaan* and *RTAL_D3OAH*, as is shown in Table 6.5. Nevertheless, the general trend was that the top influential bots on the friends list had no direct relation to the stance topic. This can be seen in the most influential follower accounts in predicting an against stance toward climate change. One of the top accounts was *AllAmericanGirl*, which had no direct relation to climate change, as is shown in Table 6.5.

6.8.3 Bots' link to stance based on the interactions type

The third research question was concerned with whether users were influenced by being exposed to posts from the social bots. We extended the effort of previous research in this field by looking beyond the bots diffusion and analyzing bot interplay with the online stances using two kinds of networks: direct interactions with bots (IN) and indirect exposure to their content (EXP). Overall, we found that users' stances were more related to bots whose content they were exposed to by following them than by directly interacting with them through retweets, replies, or mentions. This finding supports the

third hypothesis (H3) where bots shown to have presence in the direct and indirect interactions, which shows that they can affect stances even indirectly by having the user get exposed to their posted content without the need to interact directly with them.

Furthermore, we found that users with an against stance towards a given topic tended to have more indirect interactions with bots, compared to direct interactions related to the same stance toward a topic. This kind of online behavior places an emphasis on the potential hidden effect of bots, which contrasts with the existing norms of studying the effect of social bots by solely focusing on the direct interactions of users with the bot content "retweets."

6.8.4 Implications

Prior to this study, literature has informed us that bots are present in social media, and they have an effect on drifting discussions and spreading certain information related to a given topic. However, it was not clear if their presence have any relation with users' stance online. Our main findings suggested that bots' presence is linked to stance as it can be correlated with the main signals that can predict a given stance. However, our analysis shows that bots role is minimal compared to influential and famous human accounts. This core finding of our study suggests that the large fear of bots spreading messages on social media might be overrated. We do not deny the effect of their presence on the stances of people on a given topic, but we show that it is marginal compared to other factors. Our findings in this study set the path for the research community with future research opportunities to further examine the clear impact of bots on people stances by conducting qualitative studies.

Another implication should be geared towards implementing the policy of social media platforms, such as Twitter, when dealing with these accounts. It is important to increase the awareness of social media users about the effect of bots. As it has been shown, having users exposed to bots content through following them is enough to predict their stance, even more than when users interact directly with bots content through retweeting or commenting.

Finally, stance detection on social media can enable a thorough understanding of the interplay between stance towards a topic and the online signals. The ability to further analyze the hidden effect of bots as indirect interactions presents new territory for the current study of automated accounts amplification of fake news towards a certain stance (supporting/ against). The focus of these studies needs to further address the

indirect interactions instead of solo dependence on direct interactions as a retweet.

6.8.5 Limitations

Understanding bots' effect on social media is one of the highly valued questions in the social computing community (Abokhodair, Yoo, and McDonald, 2015; Zheng et al., 2019; Seering et al., 2018). However, it is challenging to study this kind of effect on the online users' stance. In our study, we used stance detection as the mean to link bots presence on users' stances by inspecting if those bots can act as predictive features to the stance. However, one limitations of our approach is that it is hard to confirm that detecting predictive bots for users' stance means that the stance has been affected by the bot not the other way around, that they interact/follow those bots since they have this stance and those bots reinforce their leaning. This is the very typical "correlation does not mean causality" problem (DeMarie-Dreblow, 1991). This is a common limitation even in existing studies that identify the bots' effect by analyzing their spread within OSNs (Aiello et al., 2012; Schuchard et al., 2019). Nevertheless, either their effect is by shaping users' stance or by reinforcing an existing stance, both still show that bots do have some role in link with polarised stances to the level that they can become predictive signals for a given stance.

Another well-known limitation on studying bots behaviour in the social network is the deleted accounts in the collected dataset. In our work we tried to address this limitation by inspecting some of those accounts manually. However, our addressing to the problem has its other limitations by itself, since we decided an account to be bot or not based on limited signals from the tweets mentioning them in the users' timelines rather than having a proper analysis of their profiles (that do not exist anymore). Unfortunately, this will remain an issue that is difficult to resolve. Nevertheless, we hope that our manual inspection of the deleted accounts gives some indication about these accounts overall behaviour.

6.9 Summary

In this study, we sought to understand the contemporary debate—admittedly, bots are everywhere, but what is the role that bots play related to polarized stances? We investigated the fourth research question of this thesis "RQ4", by examining two kinds of online user interactions: direct interactions and indirect exposure. For the direct

interactions, we evaluated users' interactions with bots with the use of mentions. As it related to indirect exposure, the analysis was carried out on the friends set of users to examine their exposure to bot content. We used the gold standard of annotated stance data that contains seven topics covering politics, religion, and social aspects. We showed empirical evidence of the effect of social bots on specific stances by using the state-of-the-art stance detection model.

Our findings indicate that users on social media tended to have limited direct interactions with social bots, that famous users in terms of followers had a sizable relationship with these stances, and that ultra-famous users tended to have the most presence on the stance interactions of specific topics from various domains. Moreover, social media users had indirect exposure to bots compared to direct interaction, which suggests that users are more exposed to bot content in an indirect manner by following these accounts, compared to direct interaction by retweets or mentions. These findings help to extend the understanding of the effect of bots on stances on social networks.

Chapter 7

Conclusion

This thesis provides a comprehensive understanding of stance detection on social media. First, it covers the literature on stance detection on social media and provides an overview of the currently available approaches for handling stance modelling. Next, an extensive analysis of the online interactions and their contribution to the current stance detection model is conducted. Then, a stance obfuscation method is designed to provide a better privacy preservation method for social media users. Finally, we show the interplay between stance and bots on social media to understand the contemporary debate on the presence of bots on the stance. The next sections summarize each chapter's findings, reiterate the key points of their work and illustrate the main contributions of the study.

7.1 Thesis contributions and findings

The research reported in this thesis has examined stance detection on social media. It has evaluated stance modelling using online factors derived from social media and assessed the overall implications. The main contributions of this thesis can be summarized as follows:

In Chapter 2, we presented a study that provided, for the first time, a comprehensive survey of stance detection on social media concerning various research fields, which mainly cover natural language processing, social computing and web science. We started by examining the current algorithms and modelling stance detection on social media. First, we defined stance and related it to the socio-linguistic domain and then extended the definition to the social media domain. We provided a taxonomy of

stance detection based on the following target types: single target, multi-related targets and claim-based target. In addition, we conducted a comparative analysis of the current modelling of stance using different algorithms by showing the performance of these models using three possible modellings (content, network and both). The use of content features was supported by modelling the stance as a textual entailment task where, for a given post, it entails an in-favor or against stance towards a target. The use of network features is based on the hypothesis of homophily on social media platforms, where individuals tend to associate and interact with similar others. We also showed how the combination of both features as multi-modelling of stance enhanced the overall performance of the stance detection model. Furthermore, we navigated the current territories of stance detection application for socio-political analysis and other veracity checking tasks. The work in this chapter has been published at IPM (Aldayel and Magdy, 2021).

In Chapter 3, we assessed the relation between stance and sentiment by testing the hypothesis and modelling of social media viewpoints concerning the polarity of sentiment and the expressed stance. To answer the first research question '*RQ1*', we extended a well-known stance dataset (Mohammad et al., 2016b) by providing a new set of tweets labelled with sentiment and stance for four topics covering various domains. This study aimed to analyse the relation between sentiment polarity and the expressed stance. We assessed the current definitions of aspect-based sentiment analysis and stance detection concerning opinion mining. We showed the relation between sentiment and stance and tested the hypothesis of modelling the social media viewpoints using sentiment polarity. The findings of this study demonstrated the weak relation between sentiment polarity and stance. They illustrate that instead of simply conjoining the sentiment and target variables as input to infer the stance, it is crucial to discriminate sentiment polarity from stance (published at SocInfo 2019 (Aldayel and Magdy, 2019a)).

Chapter 4 To answer the second research question '*RQ2*', we examined stance detection modelling by considering the content of tweets along with network features. Particularly, we assessed various online factors on social media to predict the stance on social media. In this study, we used four types of social interactions demonstrated by three sets of networks, along with the content of users' posts. Particularly, we showed a combination of different network features along with the textual content of a post,

for the first time, by using a well-known stance detection dataset (SemEval stance dataset). The study showed the most influential features for predicting the stance towards different topics. A major finding was that the influential features tend to have no direct relation with the topic of stance (topic agnostic). Another finding demonstrated the effectiveness of network features for stance detection in combination with textual content. This finding will facilitate the future stance detection studies on multi-model stance by using content along with network interactions. This study was published at CSCW 2019 (Aldayel and Magdy, 2019b).

Chapter 5 To answer the third research question '*RQ3*', we introduced a framework to preserve the privacy of social media users considering stance in social media, which contributes to the extant literature of preserving social media users' social identity. This research adds to what is known about obfuscating social identity on social media, such as gender Reddy and Knight (2016). Specifically, it provides an experimental examination of the previously held view that social media users need to control their data. This study provides a framework that enables preserving stance identity of social media users by using two techniques: addition and removal of online signals. We evaluated several stance detection algorithms on the SemEval dataset. We found that DE is more effective for obfuscation than the DR technique. This study adds to the current understanding of preserving the social identity on social media by extending the work to another kind of identity 'the online users' stance'. This work is currently under submission.

Chapter 6 We introduced a methodology to use the stance detection approach into a new field. The fake news and manipulation of viewpoints by automated accounts (bots) on social media have been studied with a focus on the spread of messages by retweets (direct interactions). In contrast, this study showed how probing the stance detection model can help with understanding the effect of these bots on stance communications. We used two type of network interactions and the stance detection model as an instrument to answer the fourth research question '*RQ4*', by measuring the relation between favor/against stance and bots. We conducted a comprehensive analysis of two types of interactions with bots: direct interactions (IN) and indirect exposure (EXP). We used the current state-of-the-art model on the SemEval dataset and extended the study by using data that have two recent topics (event datasets). We obtained new research findings that demonstrated the co-relation between bots and stance in various stance communications. Another finding highlighted the bots' noticeable presence in stance communications, which can be noticed in indirect interactions and not direct

interactions. One of the findings showed the dominance of non-bot communications in comparison to other communications (this work is under submission).

7.2 Limitations and future directions

Stance detection on social media has attracted different research communities. Various methods have been developed for a better stance modelling on social media. In this study, we assessed different methods for stance detection on social media and compared the stance with sentiment analysis. Moreover, we focused on the implication of stance detection on social media and provided a framework for mitigating the effect. Moreover, we introduced the use of stance detection in new research fields to address the effect of bots. In summary, stance detection on social media has yielded promising enhancements.

One of the limitations of this study is the scarcity of annotated data sources for the textual entailment of the stance detection task, as explained in Chapter 4, Section 2.7. Such datasets can facilitate the future directions for using stance detection for certain downstream tasks such as hate speech, bias detection, socio-political applications and rumour verification. This further intensifies the need to create an annotated stance dataset that covers low resource languages and to better understand stance modelling throughout different communities instead of solely focusing on uni-dimensional stance modelling.

Stance detection methods. Transfer learning methods have proven effective for stance detection. Transfer learning using textual content is currently the best model for stance prediction (Ghosh et al., 2019).

Using textual content is effective as the current state of stance detection focuses on single topic prediction. For instance, using the data augmentation technique to enhance the class representation for the stance detection model needs to be examined. However, cross-target/topic stance detection has been previously studied (Zhang et al., 2020). Some stance detection studies have started incorporating different types of language (cross-lingual) detection for similar and matching global events (Lai et al., 2020b; Zotova, Agerri, and Rigau, 2021). As (Kockelman, 2004) argued, stance can be interpreted as the intersection of a cross-linguistic account of the evaluated events and community-specific understandings of an individual's contribution. Providing such resources will facilitate opportunities for stance detection studies. Particularly, this can be used to address the following questions: *'How do people change language*

when stating stance towards a topic’? ’Are social attributes related to stance strongly tied to multilingual choices in comparison with cultural attitudes’? Moreover, there is currently a direction on further testing social attributes and network interactions on social media in conjunction with text as a multi-modelling technique (Lynn et al., 2019; Alkhalifa and Zubiaga, 2020). Another promising trend in stance detection studies is the incorporation of linguistic features other than sentiment polarity—such as sarcasm, which can enhance the reliability of stance detection predictions (Allaway and McKeown, 2020).

Analytical studies. Using an unsupervised algorithm can provide a robust detection of polarized stance on social media (Darwish, Magdy, and Zanoluda, 2017b; Darwish et al., 2020a; Dias and Becker, 2016). Such a stance detection method can facilitate stance analysis studies. Another possible direction for analytical studies is to address topic involvement concerning specific stances over time. Considering this fact, one of the main findings of our study (Chapter 4) is that stance detection is a topic agnostic task. This research direction has been attempted to some studies, such as (Graells-Garrido, Baeza-Yates, and Lalmas, 2020; Dong et al., 2017), where the authors in (Graells-Garrido, Baeza-Yates, and Lalmas, 2020) analysed the change in stance towards ‘abortion’ during 2008-2015. This study used a set of specific keywords to label the abortion discussion on Twitter for a given period. However, the previous approach had to consider the possibility of overlap between single-user stances along with how the stances tended to be topic agnostic, which might urge the need to further examine the evolvement on user level and analyse the expression variation of the stance. This method has been previously explored in (Borge-Holthoefer et al., 2015), which showed that the stance change over time is limited.

Stance detection for downstream tasks. An emerging research direction has started using stance for downstream tasks, such as hate-speech detection, veracity of rumour detection and bias detection.

For hate-speech detection, these studies currently focus on detecting the sub-types of hate-speech that indicate an identity-directed or indirect abuse. For instance, the authors in (Vidgen et al., 2021) designed a dataset with multiple hate-speech labels, including identity-directed abuse. In this category, the content expresses a negative statement made against an identity. Usually, the target of hateful content is a predefined competent of the expression, which can be religion, race, ethnicity, gender, sexuality, nationality, ableness or class. Another study by (Wiegand, Ruppenhofer, and Eder, 2021) addressed the detection of implicit abuse, which expressed an against iden-

tity using a non-negative sentiment by using figurative languages, such as ironies and metaphors. Some recent studies have analysed the relation between stance/partisanship and hate speech (Zannettou et al., 2020; Roussos and Dovidio, 2018). For instance, the partisanship of news articles can help analyse the magnitude of hate speech on social media platforms (Zannettou et al., 2020). Some previous studies have examined social media users' stances and hate speech towards a topic, such as (Roussos and Dovidio, 2018). One possible direction of this research is to enhance the current state of identifying identity-directed or indirect abuse by further understanding the role of stance as an auxiliary task to determine the type of abuse towards a target as an implicate or explicit abuse.

Many previous studies have used stance detection as a first step towards the veracity detection of rumours on social media. The incorporation of stance helps in debunking rumours by using social media users' stances, supporting or denying a given claim. For instance, the authors in (Li, Zhang, and Si, 2019) showed that the stance on a claim using multitask learning can help enhance the overall performance of veracity detection of rumours.

Another emerging research direction has used stance to identify bias and the partisanship of news sites (Baly et al., 2018). Identifying the partisanship of a news article can enhance fake-news detection (Baly et al., 2018). Some recent attempts have been made to use stance for probing social annotated data and social language models (Dhamala et al., 2021).

Overall, the benefit of stance detection for various downstream tasks has been proven. The current stance detection methods can be further improved by studying the cross-linguality and multi-modality of stance in social media. Moreover, the current research on stance detection needs to study the discourse structure to analyse the effectiveness of the context for stance classification on social media.

Bibliography

- Abokhodair, N.; Yoo, D.; and McDonald, D. W. 2015. Dissecting a social botnet: Growth, content and influence in twitter. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work Social Computing, CSCW '15*, 839–851. New York, NY, USA: Association for Computing Machinery.
- Abu-El-Rub, N., and Mueen, A. 2019. Botcamp: Bot-driven interactions in social campaigns. In Liu, L.; White, R. W.; Mantrach, A.; Silvestri, F.; McAuley, J. J.; Baeza-Yates, R.; and Zia, L., eds., *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, 2529–2535. ACM.
- Achananuparp, P.; Hu, X.; and Shen, X. 2008. The evaluation of sentence similarity measures. In *International Conference on data warehousing and knowledge discovery*, 305–316. Springer.
- Agarwal, A.; Singh, R.; and Toshniwal, D. 2018. Geospatial sentiment analysis using twitter data for uk-eu referendum. *Journal of Information and Optimization Sciences* 39(1):303–317.
- Aiello, L. M.; Deplano, M.; Schifanella, R.; and Ruffo, G. 2012. People are strange when you're a stranger: Impact and influence of bots on social networks. In *Sixth International AAAI Conference on Weblogs and Social Media*.
- Aker, A.; Zubiaga, A.; Bontcheva, K.; Kolliakou, A.; Procter, R.; and Liakata, M. 2017. Stance Classification in Out-of-Domain Rumours: A Case Study Around Mental Health Disorders. In *International Conference on Social Informatics*, 53–64. Springer.
- Aker, A.; Derczynski, L.; and Bontcheva, K. 2017. Simple open stance classification for rumour analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, 31–39. Varna, Bulgaria: INCOMA Ltd.

- Al-Ayyoub, M.; Rabab'ah, A.; Jararweh, Y.; Al-Kabi, M. N.; and Gupta, B. B. 2018. Studying the controversy in online crowds' interactions. *Applied Soft Computing* 66:557–563.
- Al Zamal, F.; Liu, W.; and Ruths, D. 2012. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. *ICWSM* 270:2012.
- Aldayel, A., and Magdy, W. 2019a. Assessing sentiment of the expressed stance on social media. In *Social Informatics*, 277–286. Springer International Publishing.
- Aldayel, A., and Magdy, W. 2019b. Your stance is exposed! analysing possible factors for stance detection on social media. *Proc. ACM Hum.-Comput. Interact.* 3(CSCW).
- Aldayel, A., and Magdy, W. 2021. Stance detection on social media: State of the art and trends. *Information Processing and Management* 58(4):102597.
- Alkhalifa, R., and Zubiaga, A. 2020. QMUL-SDS @ sardistance: Leveraging network interactions to boost performance on stance detection using knowledge graphs (short paper). 2765.
- Allaway, E., and McKeown, K. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8913–8931. Online: Association for Computational Linguistics.
- Allcott, H., and Gentzkow, M. 2017. Social Media and Fake News in the 2016 Election. Technical Report 2.
- An, J.; Kwak, H.; Posegga, O.; and Jungherr, A. 2019. Political discussions in homogeneous and cross-cutting communication spaces. volume 13, 68–79.
- Anand, P.; Walker, M.; Abbott, R.; Fox Tree, J. E.; Bowmani, R.; and Minor, M. 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 (WASSA 2.011)*, 1–9. Portland, Oregon: Association for Computational Linguistics.
- Augenstein, I.; Rocktäschel, T.; Vlachos, A.; and Bontcheva, K. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 876–885. Austin, Texas: Association for Computational Linguistics.

- Augenstein, I.; Vlachos, A.; and Bontcheva, K. 2016. 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. San Diego, California: Association for Computational Linguistics.
- Avnit, A. 2009. The million followers fallacy. *Pravda Media Group*.
- Baly, R.; Karadzhov, G.; Alexandrov, D.; Glass, J.; and Nakov, P. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3528–3539. Brussels, Belgium: Association for Computational Linguistics.
- Bar-Haim, R.; Bhattacharya, I.; Dinuzzo, F.; Saha, A.; and Slonim, N. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 251–261. Valencia, Spain: Association for Computational Linguistics.
- Barberá, P.; Jost, J. T.; Nagler, J.; Tucker, J. A.; and Bonneau, R. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science* 26(10):1531–1542.
- Bartlett, J., and Norrie, R. 2015. Immigration on twitter: understanding public attitudes online. *Demos*.
- Bassiouny, R. 2015. *Stance-Taking*. The International Encyclopedia of Language and Social Interaction. 1–11.
- Bastos, M. T., and Mercea, D. 2019. The brexit botnet and user-generated hyperpartisan news. *Social Science Computer Review* 37(1):38–54.
- Beigman Klebanov, B.; Beigman, E.; and Diermeier, D. 2010. Vocabulary choice as an indicator of perspective. In *Proceedings of the ACL 2010 Conference Short Papers*, 253–257. Uppsala, Sweden: Association for Computational Linguistics.
- Belkaroui, R.; Faiz, R.; and Elkhelifi, A. 2014. Conversation analysis on social networking sites. In *2014 Tenth International Conference on Signal-Image Technology and Internet-Based Systems*, 172–178. IEEE.

- Benton, A., and Dredze, M. 2018. Using author embeddings to improve tweet stance classification. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, 184–194. Brussels, Belgium: Association for Computational Linguistics.
- Bernstein, M. S.; Bakshy, E.; Burke, M.; and Karrer, B. 2013. Quantifying the invisible audience in social networks. In Mackay, W. E.; Brewster, S. A.; and Bødker, S., eds., *2013 ACM SIGCHI Conference on Human Factors in Computing Systems, CHI '13, Paris, France, April 27 - May 2, 2013*, 21–30. ACM.
- Bessi, A., and Ferrara, E. 2016. Social bots distort the 2016 us presidential election online discussion. *First Monday* 21(11-7).
- Bessi, A.; Petroni, F.; Del Vicario, M.; Zollo, F.; Anagnostopoulos, A.; Scala, A.; Caldarelli, G.; and Quattrociocchi, W. 2016. Homophily and polarization in the age of misinformation. *The European Physical Journal Special Topics* 225(10):2047–2059.
- Biber, D., and Finegan, E. 1988. Adverbial stance types in english. *Discourse Processes* 11(1):1–34.
- Blodgett, S. L.; Green, L.; and O'Connor, B. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1119–1130. Austin, Texas: Association for Computational Linguistics.
- Boichak, O.; Jackson, S.; Hemsley, J.; and Tanupabrungsun, S. 2018. Automated diffusion? bots and their influence during the 2016 us presidential election. In *International conference on information*, 17–26. Springer.
- Borge-Holthoefer, J.; Magdy, W.; Darwish, K.; and Weber, I. 2015. Content and network dynamics behind egyptian political polarization on twitter. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 700–711. Acm.
- Borges, L.; Martins, B.; and Calado, P. 2019. Combining similarity features and deep representation learning for stance detection in the context of checking fake news. *Journal of Data and Information Quality (JDIQ)* 1–26.

- Broniatowski, D. A.; Jamison, A. M.; Qi, S.; AlKulaib, L.; Chen, T.; Benton, A.; Quinn, S. C.; and Dredze, M. 2018. Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate. *American journal of public health* 108(10):1378–1384.
- Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, K. 2010. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 4.
- Cignarella, A. T.; Lai, M.; Bosco, C.; Patti, V.; Paolo, R.; et al. 2020. Sardistance evalita2020: Overview of the task on stance detection in italian tweets. In *EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, 1–10. Ceur.
- Conforti, C.; Berndt, J.; Pilehvar, M. T.; Giannitsarou, C.; Toxvaerd, F.; and Collier, N. 2020. Will-they-won't-they: A very large dataset for stance detection on Twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1715–1724. Online: Association for Computational Linguistics.
- Cossu, J.-V.; Labatut, V.; and Dugué, N. 2016. A review of features for the discrimination of twitter users: Application to the prediction of offline influence. *Social Network Analysis and Mining* 6(1):25.
- Cramér, H. 1999. *Mathematical methods of statistics (PMS-9)*. Princeton university press.
- Darwish, K.; Magdy, W.; Rahimi, A.; Baldwin, T.; and Abokhodair, N. 2018. Predicting online islamophobic behavior after parisattacks. *The Journal of Web Science* 4(3):34–52.
- Darwish, K.; Stefanov, P.; Aupetit, M.; and Nakov, P. 2020a. Unsupervised user stance detection on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 141–152.
- Darwish, K.; Stefanov, P.; Aupetit, M.; and Nakov, P. 2020b. Unsupervised user stance detection on twitter. 14:141–152.
- Darwish, K.; Magdy, W.; and Zanoouda, T. 2017a. Improved stance prediction in a user similarity feature space. In Diesner, J.; Ferrari, E.; and Xu, G., eds., *Proceedings*

- of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, Sydney, Australia, July 31 - August 03, 2017*, 145–148. ACM.
- Darwish, K.; Magdy, W.; and Zanouda, T. 2017b. Trump vs. hillary: What went viral during the 2016 us presidential election. In *Social Informatics*, 143–161. Cham: Springer International Publishing.
- Davis, C. A.; Varol, O.; Ferrara, E.; Flammini, A.; and Menczer, F. 2016. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, 273–274. International World Wide Web Conferences Steering Committee.
- DeMarie-Dreblow, D. 1991. Relation between knowledge and memory: A reminder that correlation does not imply causality. *Child Development* 62(3):484–498.
- Demszky, D.; Garg, N.; Voigt, R.; Zou, J.; Shapiro, J.; Gentzkow, M.; and Jurafsky, D. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. 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)*, 2970–3005. Minneapolis, Minnesota: Association for Computational Linguistics.
- Derczynski, L.; Bontcheva, K.; Lukasik, M.; Declerck, T.; Scharl, A.; Georgiev, G.; Osenova, P.; Lobo, T. P.; Kolliakou, A.; Stewart, R.; et al. 2015. PHEME computing veracity the fourth challenge of big social data. In *Proceedings of the Extended Semantic Web Conference EU Project Networking session ESCWPN*.
- Derczynski, L.; Bontcheva, K.; Liakata, M.; Procter, R.; Wong Sak Hoi, G.; and Zubiaga, A. 2017a. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 69–76. Vancouver, Canada: Association for Computational Linguistics.
- Derczynski, L.; Bontcheva, K.; Liakata, M.; Procter, R.; Wong Sak Hoi, G.; and Zubiaga, A. 2017b. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 69–76. Vancouver, Canada: Association for Computational Linguistics.

- Dey, K.; Shrivastava, R.; Kaushik, S.; and Mathur, V. 2017. Assessing the effects of social familiarity and stance similarity in interaction dynamics. In *International Workshop on Complex Networks and their Applications*, 843–855. Springer.
- Dey, K.; Shrivastava, R.; and Kaushik, S. 2018. Topical stance detection for twitter: A two-phase lstm model using attention. In *European Conference on Information Retrieval*, 529–536. Springer.
- Dhamala, J.; Sun, T.; Kumar, V.; Krishna, S.; Pruksachatkun, Y.; Chang, K.-W.; and Gupta, R. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. 862–872.
- Dias, M., and Becker, K. 2016. INF-UFRGS-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. San Diego, California: Association for Computational Linguistics.
- Dodds, P. S.; Harris, K. D.; Kloumann, I. M.; Bliss, C. A.; and Danforth, C. M. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one* 6(12):e26752.
- Dong, R.; Sun, Y.; Wang, L.; Gu, Y.; and Zhong, Y. 2017. Weakly-guided user stance prediction via joint modeling of content and social interaction. In Lim, E.; Winslett, M.; Sanderson, M.; Fu, A. W.; Sun, J.; Culpepper, J. S.; Lo, E.; Ho, J. C.; Donato, D.; Agrawal, R.; Zheng, Y.; Castillo, C.; Sun, A.; Tseng, V. S.; and Li, C., eds., *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, 1249–1258. ACM.
- Dori-Hacohen, S., and Allan, J. 2015. Automated controversy detection on the web. In *Advances in Information Retrieval - 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 - April 2, 2015. Proceedings*, volume 9022 of *Lecture Notes in Computer Science*, 423–434.
- Du Bois, J. W. 2007. The stance triangle. *Stance taking in discourse: Subjectivity, evaluation, interaction* 164(3):139–182.
- Dunn, A. G.; Surian, D.; Dalmazzo, J.; Rezazadegan, D.; Steffens, M.; Dyda, A.; Leask, J.; Coiera, E.; Dey, A.; and Mandl, K. D. 2020. Limited role of bots

- in spreading vaccine-critical information among active twitter users in the united states: 2017–2019. *American Journal of Public Health* 110(S3):S319–S325.
- Ebner, S.; Wang, F.; and Van Durme, B. 2019. Bag-of-words transfer: Non-contextual techniques for multi-task learning. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, 40–46. Hong Kong, China: Association for Computational Linguistics.
- Ebrahimi, J.; Dou, D.; and Lowd, D. 2016. A joint sentiment-target-stance model for stance classification in tweets. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2656–2665. Osaka, Japan: The COLING 2016 Organizing Committee.
- Elfardy, H., and Diab, M. 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. San Diego, California: Association for Computational Linguistics.
- Elfardy, H. 2007. *Perspective Identification in Informal Text*. PhD Thesis.
- Ferrara, E. 2017. Disinformation and social bot operations in the run up to the 2017 french presidential election. *First Monday* 22(8).
- Ferrara, E. 2020. Bots, elections, and social media: A brief overview. 95–114.
- Ferreira, W., and Vlachos, A. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1163–1168. San Diego, California: Association for Computational Linguistics.
- Ferreira, W., and Vlachos, A. 2019. Incorporating label dependencies in multilabel stance detection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6350–6354. Hong Kong, China: Association for Computational Linguistics.
- Fiesler, C., and Proferes, N. 2018. Participant perceptions of twitter research ethics. *Social Media+ Society* 4(1):2056305118763366.

- Fraisier, O.; Cabanac, G.; Pitarch, Y.; Besancon, R.; and Boughanem, M. 2018. Stance classification through proximity-based community detection. In *Proceedings of the 29th on Hypertext and Social Media*, Ht18, 220–228. New York: Association for Computing Machinery.
- Fuchs, C. 2018. *Social media: A critical introduction*. SAGE Publications Sage UK: London, England.
- Garimella, K., et al. 2018. *Polarization on Social Media*. Ph.D. Dissertation.
- Garimella, K., and West, R. 2019. Hot streaks on social media. *Proceedings of the International AAAI Conference on Web and Social Media* 13(01):170–180.
- Garimella, K.; De Francisc iMorales, G.; Gionis, A.; and Mathioudakis, M. 2017. Mary, mary, quite contrary: Exposing twitter users to contrarian news. In *Proceedings of the 26th International Conference on World Wide Web Companion*, 201–205. International World Wide Web Conferences Steering Committee.
- Gautam, A.; Mathur, P.; Gosangi, R.; Mahata, D.; Sawhney, R.; and Shah, R. R. 2019. # metooma: Multi-aspect annotations of tweets related to the metoo movement. *Proceedings of the International AAAI Conference on Web and Social Media* 209–216.
- Ghanem, B.; Rosso, P.; and Rangel, F. 2018. Stance detection in fake news a combined feature representation. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 66–71. Brussels, Belgium: Association for Computational Linguistics.
- Ghosh, S.; Singhanian, P.; Singh, S.; Rudra, K.; and Ghosh, S. 2019. Stance detection in web and social media: A comparative study. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, 75–87. Springer International Publishing.
- Gilani, Z.; Farahbakhsh, R.; Tyson, G.; and Crowcroft, J. 2019. A large-scale behavioural analysis of bots and humans on twitter. *ACM Trans. Web* 13(1).
- Giorgioni, S.; Politi, M.; Salman, S.; Basili, R.; and Croce, D. 2020. UNITOR @ sardistance2020: Combining transformer-based architectures and transfer learning for robust stance detection. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Work-*

- shop (EVALITA 2020)*, Online event, December 17th, 2020, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Gomaa, W. H., and Fahmy, A. A. 2013. A survey of text similarity approaches. *International Journal of Computer Applications* 68(13):13–18.
- Gong, W.; Lim, E.-P.; and Zhu, F. 2015. Characterizing silent users in social media communities. In *ICWSM*, 140–149.
- Gottipati, S.; Qiu, M.; Yang, L.; Zhu, F.; and Jiang, J. 2013. Predicting users political party using ideological stances. In *Social Informatics*, 177–191. Cham: Springer International Publishing.
- Graells-Garrido, E.; Baeza-Yates, R.; and Lalmas, M. 2020. Every colour you are: Stance prediction and turnaround in controversial issues. In *12th ACM Conference on Web Science, WebSci '20*, 174–183. New York, NY, USA: Association for Computing Machinery.
- Grcar, M.; Cherepnalkoski, D.; Mozetic, I.; and Kralj Novak, P. 2017. Stance and influence of twitter users regarding the brexit referendum. *Computational Social Networks* 4(1):6.
- Gu, Y.; Sun, Y.; Jiang, N.; Wang, B.; and Chen, T. 2014. Topic-factorized ideal point estimation model for legislative voting network. In Macskassy, S. A.; Perlich, C.; Leskovec, J.; Wang, W.; and Ghani, R., eds., *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, 183–192. ACM.
- Gu, Y.; Chen, T.; Sun, Y.; and Wang, B. 2017. Ideology detection for twitter users via link analysis. In *Social, Cultural, and Behavioral Modeling*, 262–268. Springer International Publishing.
- Gualda, E., and Rebollo, C. 2016. The refugee crisis on twitter: A diversity of discourses at a european crossroads. *Journal of Spatial and Organizational Dynamics* 4(3):199–212.
- Hamidian, S., and Diab, M. 2015. Rumor detection and classification for twitter data. In *Proceedings of the Fifth International Conference on Social Media Technologies, Communication, and Informatics*, 71–77.

- Hanawa, K.; Sasaki, A.; Okazaki, N.; and Inui, K. 2019. Stance detection attending external knowledge from wikipedia. *Journal of Information Processing* 27(1):499–506.
- Hardjono, T.; Shrier, D. L.; and Pentland, A. 2016. *Trusted Data: A New Framework for Identity and Data Sharing*. MIT Press.
- Hegelich, S., and Janetzko, D. 2016. Are social bots on twitter political actors? empirical evidence from a ukrainian social botnet. In *Tenth International AAI Conference on Web and Social Media*.
- Himmelboim, I.; McCreery, S.; and Smith, M. A. 2013. Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on twitter. *J. Comput. Mediat. Commun.* 18(2):40–60.
- Howard, P. N., and Kollanyi, B. 2016. Bots, #strongerin, and #brexit: computational propaganda during the uk-eu referendum. Available at SSRN 2798311.
- Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04*, 168–177. New York, NY, USA: Association for Computing Machinery.
- Hu, Y.; Wang, F.; and Kambhampati, S. 2013. Listening to the crowd: Automated analysis of events via aggregated twitter sentiment. In Rossi, F., ed., *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, 2640–2646. IJCAI/AAAI.
- Huber, G. A., and Malhotra, N. 2017. Political homophily in social relationships: Evidence from online dating behavior. *The Journal of Politics* 79(1):269–283.
- Hutto, C. J., and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAI conference on weblogs and social media*.
- Igarashi, Y.; Komatsu, H.; Kobayashi, S.; Okazaki, N.; and Inui, K. 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. San Diego, California: Association for Computational Linguistics.

- Jaffe, A. 2009. *Stance: sociolinguistic perspectives*. Oup Usa.
- Jakaza, E. 2020. Identity construction or obfuscation on social media: a case of Facebook and WhatsApp. *African Identities* 0(0):1–23. Publisher: Routledge _eprint: <https://doi.org/10.1080/14725843.2020.1804829>.
- Jang, M., and Allan, J. 2018. Explaining controversy on social media via stance summarization. In Collins-Thompson, K.; Mei, Q.; Davison, B. D.; Liu, Y.; and Yilmaz, E., eds., *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, 1221–1224. ACM.
- Joshi, A.; Bhattacharyya, P.; and Carman, M. 2016. Political issue extraction model: A novel hierarchical topic model that uses tweets by political and non-political authors. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 82–90. San Diego, California: Association for Computational Linguistics.
- Jurafsky, D., and Martin, J. H. 2008. *Speech and language processing*. Pearson London.
- Karamibekr, M., and Ghorbani, A. A. 2012. Sentiment analysis of social issues. In *2012 International Conference on Social Informatics*, 215–221.
- Kawintiranon, K., and Singh, L. 2021. Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4725–4735. Online: Association for Computational Linguistics.
- Kochkina, E.; Liakata, M.; and Augenstein, I. 2017. Turing at SemEval-2017 task 8: Sequential approach to rumour stance classification with branch-LSTM. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 475–480. Vancouver, Canada: Association for Computational Linguistics.
- Kockelman, P. 2004. Stance and subjectivity. *Journal of Linguistic Anthropology* 14(2):127–150.
- Kouvela, M.; Dimitriadis, I.; and Vakali, A. 2020. Bot-detective: An explainable twitter bot detection service with crowdsourcing functionalities. In *Proceedings of*

- the 12th International Conference on Management of Digital EcoSystems, MEDES '20*, 55–63. New York, NY, USA: Association for Computing Machinery.
- Krejzl, P., and Steinberger, J. 2016. UWB at SemEval-2016 task 6: Stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 408–412. San Diego, California: Association for Computational Linguistics.
- Küçük, D., and Can, F. 2020. Stance detection: A survey. *ACM Comput. Surv.* 53(1).
- Lahoti, P.; Garimella, K.; and Gionis, A. 2018. Joint non-negative matrix factorization for learning ideological leaning on twitter. In Chang, Y.; Zhai, C.; Liu, Y.; and Maarek, Y., eds., *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, 351–359. ACM.
- Lai, M.; Farías, D. I. H.; Patti, V.; and Rosso, P. 2016. Friends and enemies of clinton and trump: using context for detecting stance in political tweets. In *Mexican International Conference on Artificial Intelligence*, 155–168. Springer.
- Lai, M.; Patti, V.; Ruffo, G.; and Rosso, P. 2018. Stance evolution and twitter interactions in an italian political debate. In *Natural Language Processing and Information Systems*, 15–27. Springer International Publishing.
- Lai, M.; Cignarella, A. T.; Farías, D. I. H.; Bosco, C.; Patti, V.; and Rosso, P. 2020a. Multilingual Stance Detection in Social Media Political Debates. *Computer Speech & Language* 101075.
- Lai, M.; Cignarella, A. T.; Hernández Farías, D. I.; Bosco, C.; Patti, V.; and Rosso, P. 2020b. Multilingual stance detection in social media political debates. *Computer Speech Language* 63:101075.
- Lee, C.; Kwak, H.; Park, H.; and Moon, S. B. 2010. Finding influentials based on the temporal order of information adoption in twitter. In Rappa, M.; Jones, P.; Freire, J.; and Chakrabarti, S., eds., *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, 1137–1138. ACM.
- Lee, H. W. 2018. Using twitter hashtags to gauge real-time changes in public opinion: An examination of the 2016 us presidential election. In *International Conference on Social Informatics*, 168–175. Springer.

- Levinson, D., and Ember, M. 1996. *Encyclopedia of cultural anthropology*. Holt New York.
- Levy, O., and Goldberg, Y. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 302–308. Baltimore, Maryland: Association for Computational Linguistics.
- Li, Y., and Caragea, C. 2019. Multi-task stance detection with sentiment and stance lexicons. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6299–6305. Hong Kong, China: Association for Computational Linguistics.
- Li, C.; Porco, A.; and Goldwasser, D. 2018. Structured representation learning for on-line debate stance prediction. In *Proceedings of the 27th International Conference on Computational Linguistics*, 3728–3739. Santa Fe, New Mexico, USA: Association for Computational Linguistics.
- Li, Q.; Zhang, Q.; and Si, L. 2019. Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1173–1179. Florence, Italy: Association for Computational Linguistics.
- Liebetrau, A. M. 1983. *Measures of association*. Sage.
- Lin, W.-H.; Wilson, T.; Wiebe, J.; and Hauptmann, A. 2006. Which side are you on? identifying perspectives at the document and sentence levels. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, 109–116. New York City: Association for Computational Linguistics.
- Liu, C.; Li, W.; Demarest, B.; Chen, Y.; Couture, S.; Dakota, D.; Haduong, N.; Kaufman, N.; Lamont, A.; Pancholi, M.; Steimel, K.; and Kübler, S. 2016. 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. San Diego, California: Association for Computational Linguistics.
- Luceri, L.; Deb, A.; Badawy, A.; and Ferrara, E. 2019. Red bots do it better: Comparative analysis of social bot partisan behavior. In *Companion Proceedings of The 2019 World Wide Web Conference*, 1007–1012. ACM.

- Lynn, V.; Giorgi, S.; Balasubramanian, N.; and Schwartz, H. A. 2019. Tweet classification without the tweet: An empirical examination of user versus document attributes. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, 18–28. Minneapolis, Minnesota: Association for Computational Linguistics.
- Ma, J.; Gao, W.; and Wong, K.-F. 2018. Detect rumor and stance jointly by neural multi-task learning. In *Companion of the The Web Conference 2018*, 585–593.
- Ma, Y.; Peng, H.; and Cambria, E. 2018. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. In McIlraith, S. A., and Weinberger, K. Q., eds., *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 5876–5883. AAAI Press.
- Magdy, W.; Darwish, K.; Abokhodair, N.; Rahimi, A.; and Baldwin, T. 2016. #isisisnotislam or# deportallmuslims?: Predicting unspoken views. In *Proceedings of the 8th ACM Conference on Web Science*, 95–106. Acm.
- McKendrick, D., and Webb, S. 2014. Taking a political stance in social work. *Critical and Radical Social Work* 2(3):357–369.
- McKnight, P. E., and Najab, J. 2010. Mann-whitney u test. *The Corsini encyclopedia of psychology* 1–1.
- Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; and Cherry, C. 2016a. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 3945–3952. Portorož, Slovenia: European Language Resources Association (ELRA).
- Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; and Cherry, C. 2016b. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 31–41. San Diego, California: Association for Computational Linguistics.
- Mohammad, S. M.; Sobhani, P.; and Kiritchenko, S. 2017. Stance and sentiment in tweets. *ACM Trans. Internet Technol (2017)*. 17(3).

- Mohammad, S. 2016. A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 174–179. San Diego, California: Association for Computational Linguistics.
- Mohtarami, M.; Baly, R.; Glass, J.; Nakov, P.; Màrquez, L.; and Moschitti, A. 2018. Automatic stance detection using end-to-end memory networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 767–776. New Orleans, Louisiana: Association for Computational Linguistics.
- Murakami, A., and Raymond, R. 2010. Support or oppose? classifying positions in online debates from reply activities and opinion expressions. In *Coling 2010: Posters*, 869–875. Beijing, China: Coling 2010 Organizing Committee.
- Murphy, J.; Hill, C. A.; and Dean, E. 2014. Social media, sociality and survey research. *Social media, sociality and survey research* 1–33.
- Nakov, P.; Rosenthal, S.; Kozareva, Z.; Stoyanov, V.; Ritter, A.; and Wilson, T. 2013a. SemEval-2013 task 2: Sentiment analysis in Twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 312–320. Atlanta, Georgia, USA: Association for Computational Linguistics.
- Nakov, P.; Rosenthal, S.; Kozareva, Z.; Stoyanov, V.; Ritter, A.; and Wilson, T. 2013b. SemEval-2013 task 2: Sentiment analysis in Twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 312–320. Atlanta, Georgia, USA: Association for Computational Linguistics.
- Newman, N. 2011. Mainstream media and the distribution of news in the age of social discovery (risj reports). *Reuters Institute for the Study of Journalism, University of Oxford* 1–58.
- Nyhan, B., and Reifler, J. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32(2):303–330.
- Ochs, E. 1996. *Linguistic resources for socializing humanity*. Cambridge University Press.

- Overbey, L. A.; Batson, S. C.; Lyle, J.; Williams, C.; Regal, R.; and Williams, L. 2017. Linking twitter sentiment and event data to monitor public opinion of geopolitical developments and trends. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, 223–229. Springer.
- Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*. 2(1–2):1–135.
- Park, S.; Ko, M.; Kim, J.; Liu, Y.; and Song, J. 2011. The politics of comments: predicting political orientation of news stories with commenters’ sentiment patterns. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, 113–122. ACM.
- Patra, B. G.; Das, D.; and Bandyopadhyay, S. 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. San Diego, California: Association for Computational Linguistics.
- Patwa, P.; Aguilar, G.; Kar, S.; Pandey, S.; PYKL, S.; Garrette, D.; Gambäck, B.; Chakraborty, T.; Solorio, T.; and Das, A. 2020. Semeval-2020 sentimix task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*. Barcelona, Spain: Association for Computational Linguistics.
- Pennacchiotti, M., and Popescu, A. 2011. Democrats, republicans and starbucks aficionados: user classification in twitter. In Apté, C.; Ghosh, J.; and Smyth, P., eds., *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*, 430–438. ACM.
- Perez, B.; Musolesi, M.; and Stringhini, G. 2018. You are your metadata: Identification and obfuscation of social media users using metadata information. In *Twelfth International AAAI Conference on Web and Social Media*.
- Pontiki, M.; Galanis, D.; Pavlopoulos, J.; Papageorgiou, H.; Androutsopoulos, I.; and Manandhar, S. 2014a. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 27–35. Dublin, Ireland: Association for Computational Linguistics.

- Pontiki, M.; Galanis, D.; Pavlopoulos, J.; Papageorgiou, H.; Androutsopoulos, I.; and Manandhar, S. 2014b. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 27–35. Dublin, Ireland: Association for Computational Linguistics.
- Puertas, E.; Moreno-Sandoval, L. G.; Plaza-del Arco, F. M.; Alvarado-Valencia, J. A.; Pomares-Quimbaya, A.; and Alfonso, L. 2019. Bots and gender profiling on twitter using sociolinguistic features.
- Pulido, C. M.; Redondo-Sama, G.; Sordé-Martí, T.; and Flecha, R. 2018. Social impact in social media: A new method to evaluate the social impact of research. *PloS one* 13(8):e0203117.
- Qazvinian, V.; Rosengren, E.; Radev, D. R.; and Mei, Q. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1589–1599. Edinburgh, Scotland, UK.: Association for Computational Linguistics.
- Qiu, M.; Sim, Y.; Smith, N. A.; and Jiang, J. 2015. Modeling user arguments, interactions, and attributes for stance prediction in online debate forums. In Venkatasubramanian, S., and Ye, J., eds., *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, BC, Canada, April 30 - May 2, 2015*, 855–863. SIAM.
- Quattrociocchi, W.; Scala, A.; and Sunstein, C. R. 2016. Echo chambers on facebook. *Available at SSRN 2795110*.
- Rajadesingan, A., and Liu, H. 2014. Identifying users with opposing opinions in twitter debates. In *International conference on social computing, behavioral-cultural modeling, and prediction*, 153–160. Springer.
- Reddy, S., and Knight, K. 2016. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, 17–26. Austin, Texas: Association for Computational Linguistics.
- Rizoiu, M.-A.; Graham, T.; Zhang, R.; Zhang, Y.; Ackland, R.; and Xie, L. 2018. #debatenight: The role and influence of socialbots on twitter during the 1st 2016 us presidential debate. In *Twelfth International AAAI Conference on Web and Social Media*.

- Roussos, G., and Dovidio, J. F. 2018. Hate speech is in the eye of the beholder: The influence of racial attitudes and freedom of speech beliefs on perceptions of racially motivated threats of violence. *Social Psychological and Personality Science* 9(2):176–185.
- Santia, G. C.; Mujib, M. I.; and Williams, J. R. 2019. Detecting social bots on facebook in an information veracity context. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 463–472.
- Schmidt, D. 2015. Stock market rumors and credibility. *Document de travail, HEC Paris*.
- Schuchard, R.; Crooks, A. T.; Stefanidis, A.; and Croitoru, A. 2019. Bot stamina: examining the influence and staying power of bots in online social networks. *Applied Network Science* 4(1):55.
- Seering, J.; Flores, J. P.; Savage, S.; and Hammer, J. 2018. The Social Roles of Bots: Evaluating Impact of Bots on Discussions in Online Communities. *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW):157:1–157:29.
- Sen, A.; Sinha, M.; Mannarswamy, S.; and Roy, S. 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.
- Sen, I.; Flöck, F.; and Wagner, C. 2020. On the reliability and validity of detecting approval of political actors in tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1413–1426. Online: Association for Computational Linguistics.
- Shao, C.; Ciampaglia, G. L.; Varol, O.; Yang, K.-C.; Flammini, A.; and Menczer, F. 2018. The spread of low-credibility content by social bots. *Nature communications* 9(1):4787.
- Shu, K.; Wang, S.; and Liu, H. 2019. Beyond news contents: The role of social context for fake news detection. In Culpepper, J. S.; Moffat, A.; Bennett, P. N.; and Lerman, K., eds., *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019*, 312–320. ACM.

- Shultz, J. M.; Baingana, F.; and Neria, Y. 2014. The 2014 ebola outbreak and mental health: current status and recommended response. *Jama* 567–568.
- Siddiqua, U. A.; Chy, A. N.; and Aono, M. 2018. Stance detection on microblog focusing on syntactic tree representation. In *International Conference on Data Mining and Big Data*, 478–490. Springer.
- Siddiqua, U. A.; Chy, A. N.; and Aono, M. 2019a. Tweet stance detection using an attention based neural ensemble model. 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)*, 1868–1873. Minneapolis, Minnesota: Association for Computational Linguistics.
- Siddiqua, U. A.; Chy, A. N.; and Aono, M. 2019b. Tweet Stance Detection Using Multi-Kernel Convolution and Attentive LSTM Variants. *IEICE Transactions on Information and Systems* E102d:2493–2503.
- Simaki, V.; Paradis, C.; Skeppstedt, M.; Sahlgren, M.; Kucher, K.; and Kerren, A. 2020. Annotating speaker stance in discourse: the brexit blog corpus. *Corpus Linguistics and Linguistic Theory* 16(2):215–248.
- Simaki, V.; Paradis, C.; and Kerren, A. 2017. Stance classification in texts from blogs on the 2016 british referendum. In *Speech and Computer*, 700–709. Springer International Publishing.
- Singh, P. K.; Singh, S. K.; and Paul, S. 2015. Sentiment classification of social issues using contextual valence shifters. *International Journal of Engineering and Technology* 7(4):1443–1452.
- Smith, K. S.; McCreddie, R.; Macdonald, C.; and Ounis, I. 2017. Analyzing disproportionate reaction via comparative multilingual targeted sentiment in twitter. In Diesner, J.; Ferrari, E.; and Xu, G., eds., *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, Sydney, Australia, July 31 - August 03, 2017*, 317–320. ACM.
- Sobhani, P.; Inkpen, D.; and Zhu, X. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 551–557. Valencia, Spain: Association for Computational Linguistics.

- Sobhani, P.; Inkpen, D.; and Zhu, X. 2019. Exploring deep neural networks for multitarget stance detection. *Computational Intelligence* 35(1):82–97.
- Sobhani, P.; Mohammad, S.; and Kiritchenko, S. 2016. Detecting stance in tweets and analyzing its interaction with sentiment. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, 159–169. Berlin, Germany: Association for Computational Linguistics.
- Sobhani, P. 2017. *Stance Detection and Analysis in Social Media*. Ph.D. Dissertation, Université d’Ottawa/University of Ottawa.
- Somasundaran, S., and Wiebe, J. 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*, 226–234. Suntec, Singapore: Association for Computational Linguistics.
- Somasundaran, S., and Wiebe, J. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 116–124. Los Angeles, CA: Association for Computational Linguistics.
- Stella, M.; Ferrara, E.; and De Domenico, M. 2018. Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences* 12435–12440.
- Sun, Q.; Wang, Z.; Zhu, Q.; and Zhou, G. 2018. Stance detection with hierarchical attention network. In *Proceedings of the 27th International Conference on Computational Linguistics*, 2399–2409. Santa Fe, New Mexico, USA: Association for Computational Linguistics.
- Tang, D.; Qin, B.; and Liu, T. 2016. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 214–224. Austin, Texas: Association for Computational Linguistics.
- Taulé, M.; Pardo, F. M. R.; Martí, M. A.; and Rosso, P. 2018a. Overview of the task on multimodal stance detection in tweets on catalan #1oct referendum. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society*

- for *Natural Language Processing (SEPLN 2018)*, Sevilla, Spain, September 18th, 2018, volume 2150 of *CEUR Workshop Proceedings*, 149–166. CEUR-WS.org.
- Taulé, M.; Pardo, F. M. R.; Martí, M. A.; and Rosso, P. 2018b. Overview of the task on multimodal stance detection in tweets on catalan# 1oct referendum. In *IberEval SEPLN*, 149–166.
- Taulé, M.; Martí, M. A.; Rangel, F. M.; Rosso, P.; Bosco, C.; and Patti, V. 2017. Overview of the task on stance and gender detection in tweets on Catalan independence at IberEval 2017. In *2nd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2017*, volume 1881, 157–177. CEUR-WS.
- Thelwall, M. 2017. *The Heart and Soul of the Web? Sentiment Strength Detection in the Social Web with SentiStrength*. Cham: Springer International Publishing. 119–134.
- Thonet, T.; Cabanac, G.; Boughanem, M.; and Pinel-Sauvagnat, K. 2017. Users are known by the company they keep: Topic models for viewpoint discovery in social networks. In Lim, E.; Winslett, M.; Sanderson, M.; Fu, A. W.; Sun, J.; Culpepper, J. S.; Lo, E.; Ho, J. C.; Donato, D.; Agrawal, R.; Zheng, Y.; Castillo, C.; Sun, A.; Tseng, V. S.; and Li, C., eds., *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, 87–96. ACM.
- Trabelsi, A., and Zaïane, O. R. 2018. Unsupervised model for topic viewpoint discovery in online debates leveraging author interactions. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, 425–433. AAAI Press.
- Tsolmon, B.; Kwon, A.-R.; and Lee, K.-S. 2012. Extracting social events based on timeline and sentiment analysis in twitter corpus. In *International Conference on Application of Natural Language to Information Systems*, 265–270. Springer.
- Unankard, S.; Li, X.; Sharaf, M.; Zhong, J.; and Li, X. 2014. Predicting elections from social networks based on sub-event detection and sentiment analysis. In *International Conference on Web Information Systems Engineering*, 1–16. Springer.
- Varol, O.; Ferrara, E.; Davis, C. A.; Menczer, F.; and Flammini, A. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Eleventh international AAAI conference on web and social media*.

- Vidgen, B.; Nguyen, D.; Margetts, H.; Rossini, P.; and Tromble, R. 2021. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2289–2303. Online: Association for Computational Linguistics.
- Vijayaraghavan, P.; Sysoev, I.; Vosoughi, S.; and Roy, D. 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)*, 413–419. San Diego, California: Association for Computational Linguistics.
- Voigt, R.; Jurgens, D.; Prabhakaran, V.; Jurafsky, D.; and Tsvetkov, Y. 2018. RtGender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).
- Walker, M. A.; Anand, P.; Abbott, R.; Tree, J. E. F.; Martell, C.; and King, J. 2012. That is your evidence?: Classifying stance in online political debate. *Decision Support Systems* 719–729.
- Waniek, M.; Michalak, T. P.; Wooldridge, M. J.; and Rahwan, T. 2018. Hiding individuals and communities in a social network. *Nature Human Behaviour* 2(2):139.
- Weber, I.; Garimella, V. R. K.; and Batayneh, A. 2013. Secular vs. islamist polarization in egyp on twitter. In Rokne, J. G., and Faloutsos, C., eds., *Advances in Social Networks Analysis and Mining 2013, ASONAM '13, Niagara, ON, Canada - August 25 - 29, 2013*, 290–297. ACM.
- Wei, W.; Zhang, X.; Liu, X.; Chen, W.; and Wang, T. 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. San Diego, California: Association for Computational Linguistics.
- Wei, P.; Lin, J.; and Mao, W. 2018. Multi-target stance detection via a dynamic memory-augmented network. In Collins-Thompson, K.; Mei, Q.; Davison, B. D.; Liu, Y.; and Yilmaz, E., eds., *The 41st International ACM SIGIR Conference on*

- Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, 1229–1232. ACM.
- Wei, P.; Mao, W.; and Chen, G. 2019. A topic-aware reinforced model for weakly supervised stance detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7249–7256.
- Wiegand, M.; Ruppenhofer, J.; and Eder, E. 2021. Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 576–587. Online: Association for Computational Linguistics.
- Williams, M. L.; Burnap, P.; and Sloan, L. 2017. Towards an ethical framework for publishing twitter data in social research: Taking into account users’ views, online context and algorithmic estimation. *Sociology* 51(6):1149–1168.
- Wojatzki, M., and Zesch, T. 2016. Itl.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. San Diego, California: Association for Computational Linguistics.
- Xi, N.; Ma, D.; Liou, M.; Steinert-Threlkeld, Z. C.; Anastasopoulos, J.; and Joo, J. 2020. Understanding the political ideology of legislators from social media images. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 726–737.
- Xu, Q.; Qu, L.; Xu, C.; and Cui, R. 2019. Privacy-aware text rewriting. In *Proceedings of the 12th International Conference on Natural Language Generation*, 247–257. Tokyo, Japan: Association for Computational Linguistics.
- Yang, K.-C.; Varol, O.; Davis, C. A.; Ferrara, E.; Flammini, A.; and Menczer, F. 2019. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies* 1(1):48–61.
- Yang, D.; Qu, B.; and Cudré-Mauroux, P. 2019. Privacy-preserving social media data publishing for personalized ranking-based recommendation. *IEEE Transactions on Knowledge and Data Engineering* 31(3):507–520.

- Zannettou, S.; ElSherief, M.; Belding, E.; Nilizadeh, S.; and Stringhini, G. 2020. Measuring and characterizing hate speech on news websites. In *12th ACM Conference on Web Science*, 125–134.
- Zarrella, G., and Marsh, A. 2016. MITRE at SemEval-2016 task 6: Transfer learning for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 458–463. San Diego, California: Association for Computational Linguistics.
- Zhang, Q.; Liang, S.; Lipani, A.; Ren, Z.; and Yilmaz, E. 2019. From stances' imbalance to their hierarchical representation and detection. In Liu, L.; White, R. W.; Mantrach, A.; Silvestri, F.; McAuley, J. J.; Baeza-Yates, R.; and Zia, L., eds., *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, 2323–2332. ACM.
- Zhang, B.; Yang, M.; Li, X.; Ye, Y.; Xu, X.; and Dai, K. 2020. Enhancing cross-target stance detection with transferable semantic-emotion knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3188–3197. Online: Association for Computational Linguistics.
- Zheng, L. N.; Albano, C. M.; Vora, N. M.; Mai, F.; and Nickerson, J. V. 2019. The roles bots play in wikipedia. *Proc. ACM Hum.-Comput. Interact.* 3(CSCW).
- Zhou, Y.; Cristea, A. I.; and Shi, L. 2017. Connecting targets to tweets: Semantic attention-based model for target-specific stance detection. In *International Conference on Web Information Systems Engineering*, 18–32. Springer.
- Zhu, L.; He, Y.; and Zhou, D. 2019. Hierarchical viewpoint discovery from tweets using bayesian modelling. *Expert Systems with Applications* 116:430–438.
- Zotova, E.; Agerri, R.; and Rigau, G. 2021. Semi-automatic generation of multilingual datasets for stance detection in twitter. *Expert Systems with Applications* 170:114547.
- Zubiaga, A.; Liakata, M.; Procter, R.; Wong Sak Hoi, G.; and Tolmie, P. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS ONE* 11(3):1–29.
- Zubiaga, A.; Aker, A.; Bontcheva, K.; Liakata, M.; and Procter, R. 2017. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.*

Zubiaga, A.; Kochkina, E.; Liakata, M.; Procter, R.; Lukasik, M.; Bontcheva, K.; Cohn, T.; and Augenstein, I. 2018. Discourse-aware rumour stance classification in social media using sequential classifiers. *Information Processing & Management* 273–290.

Appendix A

Background

A.1 List of recent stance prediction and detection work

Study	Features	ML	Dataset
Darwish et al. (2018)	Content Features (Hashtags, Text); Profile Features (Description, Name, Location); Network Features (Mention, Reply, Retweet)	SVM	Islamophobic dataset (Twitter) [Not available]
Magdy et al. (2016)	Content Features (Hashtags, Text); Profile Features (Description, Name, Location); Network Features (Mention, Reply, Retweet)	SVM	Islamophobic dataset (Twitter) [Not available]
Darwish, Magdy, and Zanoouda (2017a)	Content Features(Text); Interaction Elements(retweeted accounts, used hashtags, mentioned accounts (MEN), shared URLs (URL)); User Similarity	SVM	Islands Dataset and Islamophobic dataset (Twitter) [Not available]

Lahoti, Garimella, and Gionis (2018)	A combination of network and content	Non-negative matrix factorization	dataset covered Three controversial topics:gun control,abortion and obamacare (Twitter) [Not available]
Gottipati et al. (2013)	similarity between users	Probabilistic Matrix Factorization	1000 user profile of Democrats and Republicans (debate.org) [Not available]
Dong et al. (2017)	post level interaction and user level interaction	Stance-based Text Generative Model with Rule-based User-User Interaction Links	CNN dataset, 4Forums and IAC discussion forum [Not available]

Table A.1: Work in stance prediction

Study	Task	Features	ML	Dataset
Aldayel and Magdy (2019b)	target-specific	NW features	SVM	SemEval-2016 shared task 6 [Available]
Lynn et al. (2019)	target-specific	NW (followee)	RNN	SemEval-2016 shared task 6 [Available]
Siddiqua, Chy, and Aono (2019a)	target-specific	Content	Nested LSTMs	SemEval-2016 shared task 6 [Available]
Sun et al. (2018)	target-specific	Content	Hierarchical Attention NN	SemEval-2016 shared task 6 [Available]
Siddiqua, Chy, and Aono (2018)	target-specific	Content	SVM Tree Kernel	SemEval-2016 shared task 6 [Available]

Wei, Mao, and Chen (2019)	target-specific	Content+Sentiment lexicon	BiLSTM	SemEval-2016 shared task 6 [Available]
Wei, Mao, and Chen (2019)	target-specific	Content+Noisy stance labeling + Topic Modeling	BiGRU	SemEval-2016 shared task 6 [Available]
Ebner, Wang, and Van Durme (2019)	target-specific	words embedding	Deep averaging network	SemEval-2016 shared task 6 [Available]
Liu et al. (2016)	target-specific	bag-of-words and word vectors (GloVe and word2vec)	gradient boosting decision trees and SVM and merge all classifiers into an ensemble method	SemEval-2016 shared task 6 [Available]
Dias and Becker (2016)	target-specific	n-gram and sentiment	SVM	SemEval-2016 shared task 6 [Available]
Dias and Becker (2016)	target-specific	n-gram and sentiment	SVM	SemEval-2016 shared task 6 [Available]
Igarashi et al. (2016)	target-specific	Reply, BagOfWord, BagOfDependencies, POS tags Sentiment WordNet, Sentiment Word Subject, Target Sentiment and Point-wise Mutual Information	CNN	SemEval-2016 shared task 6 [Available]

Igarashi et al. (2016)	target-specific	Reply, BagOfWord, BagOfDependencies, POS tags Sentiment WordNet, Sentiment Word Subject, Target Sentiment and Point-wise Mutual Information	CNN	SemEval-2016 shared task 6 [Available]
Augenstein et al. (2016)	target-specific	word2vec	Bidirectional LSTMs	SemEval-2016 shared task 6 [Available]
Krejzl and Steinberger (2016)	target-specific	hashtags, n-grams, tweet length, Part-of-speech, General Inquirer, entity-centered sentiment dictionaries, Domain Stance Dictionary	Maximum entropy classifier	SemEval-2016 shared task 6 [Available]
Ebrahimi, Dou, and Lowd (2016)	target-specific	n-gram and sentiments	discriminative and generative models	SemEval-2016 shared task 6 [Available]
Wei et al. (2016)	target-specific	Google news word2vec and hashtags	CNN	SemEval-2016 shared task 6 [Available]
Zarrella and Marsh (2016)	target-specific	word2vec hash-tags	LSTM	SemEval-2016 shared task 6 [Available]
Rajadesingan and Liu (2014)	target-specific	unigrams, bigrams and trigrams	Naive Bayes	hotly contested gun reforms debate from April 15th, 2013 to April 18th, 2013. [Available]

Zhou, Cristea, and Shi (2017)	target-specific	word embeddings	Bi-directional GRU-CNN	SemEval-2016 shared task 6 [Available]
Vijayaraghavan et al. (2016)	target-specific	word embeddings	convolutional neural networks(CNN)	SemEval-2016 shared task 6 [Available]
Elfardy and Diab (2016)	target-specific	Lexical Features, Latent Semantics, Sentiment, Linguistic Inquiry , Word Count and Frame Semantics features	SVM	SemEval-2016 shared task 6 [Available]
Lai et al. (2016)	target-specific	sentiment, Opinion target, Structural Features(Hashtags, Mentions, Punctuation marks), text-Based Features(targetByName, targetByPronoun, targetParty, targetPartyColleagues)	Gaussian Naive Bayes classifier	Hillary Clinton and Donald Trump dataset [Not available]
Sobhani, Inkpen, and Zhu (2017)	multi-target	word vectors	bidirectional recurrent neural network (RNN)	Multi-Target Stance dataset [Available]
Siddiqua, Chy, and Aono (2019b)	multi-target	content of tweets	Multi-Kernel Convolution and Attentive LSTM	Multi-Target Stance dataset [Available]

Bar-Haim et al. (2017)	claim-based	Contrast scores	random forest and SVM	The claim polarity dataset claims are from Wikipedia and motions f are from rom (IDEA)(on-line forums) [Available]
Aker, Derczynski, and Bontcheva (2017)	claim-based	Linguistic, message-based, and topic-based such as (Bag of words, POS tag, Sentiment, Named entity and others	Random Forest, Decision tree and Instance Based classifier (K-NN)	RumourEval and PHEME datasets [Available]
Hamidian and Diab (2015)	claim-based	tweet content, Unigram-Bigram Bag of Words, Part of Speech, Sentiment, Emoticon, Named-Entity Recognition , event, time, Reply, Re-tweet, User ID, Hashtag, URL	decision trees	Qazvinian et al. (2011) [Available]
Aker, Derczynski, and Bontcheva (2017)	claim-based	BOW,Brown Cluster, POS tag, Sentiment, Named entity, Reply, Emoticon, URL, Mood, Originality score, is User Verified(0-1),Number Of Followers, Role score, Engagement score, Favourites score and other tweets related features	decision tree, Random Forests and Instance Based classifier	RumourEval dataset Derczynski et al. (2017b) and the PHEME dataset Derczynski et al. (2015) [Available]

Zubiaga et al. (2018)	claim-based	Word2Vec, POS, Use of negation, Use of swear words, Tweet length, Word count, Use of question mark, Use of exclamation mark, Attachment of URL and other contextualized features	Linear CRF and tree CRF , a Long Short-Term Memory (LSTM)	PHEME dataset Derczynski et al. (2015) and Rmour dataset associated with eight events corresponding to breaking news events Zubiaga et al. (2016) [Available]
Kochkina, Liakata, and Augenstein (2017)	claim-based	word2vec, Tweet lexicon (count of negation words and count of swear words),Punctuation, Attachments,Relation to other tweets, Content length and Tweet role (source tweet of a conversation)	branch-LSTM , a neural network architecture that uses layers of LSTM units	Rumoureal dataset Derczynski et al. (2017b) [Available]

Table A.2: Work in stance classification

A.2 Datasets for stance classification and prediction tasks

















Dataset	Type	Lg	Topics	Stance annotations	Size
SemEval-Stance Task A Mohammad et al. (2016b)	T 	EN	Atheism, Climate, Feminism, Hillary, and Abortion	$S(T, Tweet) = \{Favor, Against, Neither\}$	4,163
SemEval-Stance Task B Mohammad et al. (2016b)	T 	En	Donald Trump	$S(T, Tweet) = \{Favor, Against, Neither\}$	707
SemEval-Rumours2017 Derczynski et al. (2017b)	C 	En	Various events	$S(C, Reply) = \{Support, Deny, Commenting\}$	5,568
Multi-Targets Stance Sobhani, Inkpen, and Zhu (2017)	MT 	En	Hillary-Sanders, Hillary-Trump and Cruz-Trump	$S(T1 T2, Tweet) = \{FavorT1 AgainstT2\}$	4,455
Trump-Hillary Darwish, Magdy, and Zanouda (2017b)	MT 	En	Hillary Clinton, Donald Trump	$S(T1 T2, Tweet) = \{FavorT1 AgainstT2\}$	3,450
MultiStanceCat Taulé et al. (2018b)	T  , 	Es, Ca	Catalan Referendum	$S(T, Tweet) = \{Favor, Against, Neither\}$	11,398
RumourEval2019 ?	C  , 	En	Various events	$S(C, Reply) = \{Support, Deny, Commenting\}$	5,216
Me Too dataset Gautam et al. (2019)	T 	En	Me Too movement	$S(T, Tweet) = \{Support, Opposition\}$	9,973
WT-WT Conforti et al. (2020)	C 	En	Companies mergers and acquisitions	$S(C, Tweet) = \{Support, Refute, Comment, Unrelated\}$	51,284
Sardi-Stance-EVALITA2020 Cignarella et al. (2020)	T  , 	It	Sardines movement	$S(T, Tweet) = \{Favor, Against, Neither\}$	3,242

Table A.3: Publicly available data-sets with stance annotations for stance classification in social media (in chronological order). Sources:  Twitter and  Reddit. Types "C": Claim-based, "T": Target based, and "MT": Multi related targets. The  indicates a multi-model dataset that contains contextual data along with the text. In stance annotation S, inputs "T": Target and "C": Claim.






Dataset	Type	Lg	Topics	Stance annotations	Size
Create-Debate Qiu et al. (2015)	Micro-level + Macro-level 	En	Various topics	The probability of user choosing a stance on an issue	39,549
User Stance dataset Dong et al. (2017)	Micro-level 	En	Online-forums	S(Issue,Posts)={Pro,Con}	1,983,222
Islands (E) and Islamophobia (I) Darwish et al. (2018)	Micro-level 	En, Ar	Egypt Islands and Islamophobia	S(Previous tweets,Users,Topic)={Pos,Neg}	40,227

Table A.4: Publicly available data-sets with stance annotations for stance predictions (in chronological order). Sources: : Twitter and : News comments or online forum data.

Appendix B

Stance obfuscation

B.1 Survey results

To evaluate the degree to which Twitter users feel the need to keep their stance private, we surveyed 1,143 participants recruited through Amazon Mechanical Turk. In order to be eligible, respondents had to be at least 18 years old, live in the US, and have a Twitter account for at least one year.

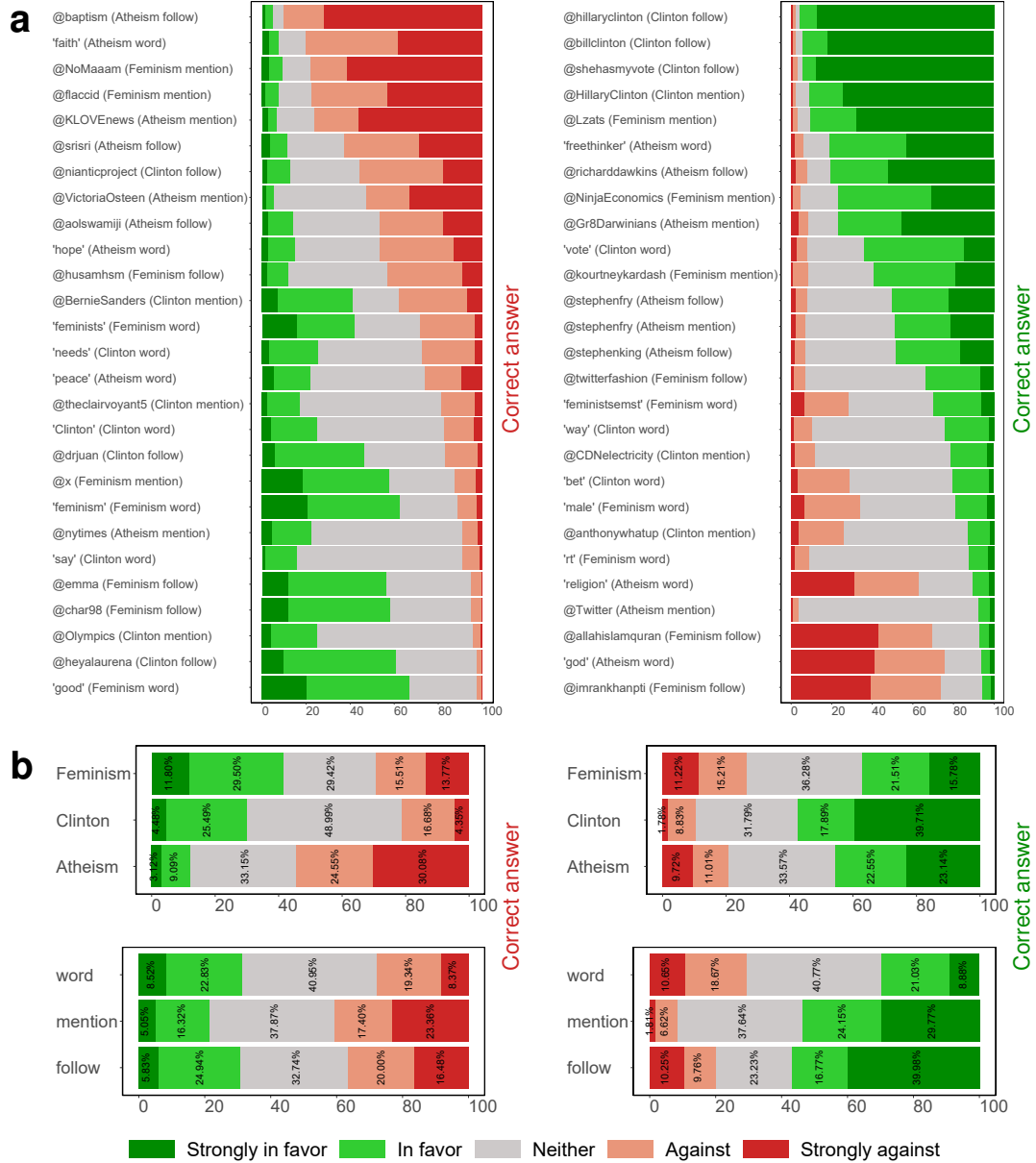


Figure B.1: Participants' ability to identify the stance indicated by different features.

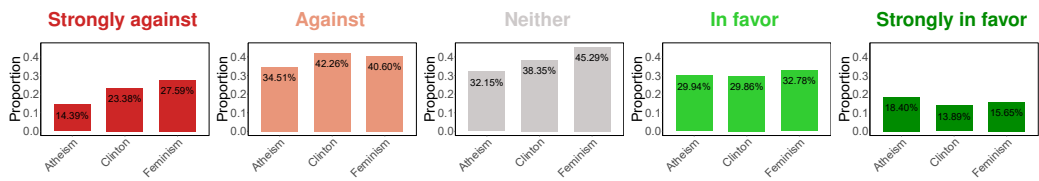


Figure B.2: The degree to which participants feel the need to avoid revealing their stance.

Appendix C

Socialbots and stance

C.1 Distribution of accounts scores on topic level

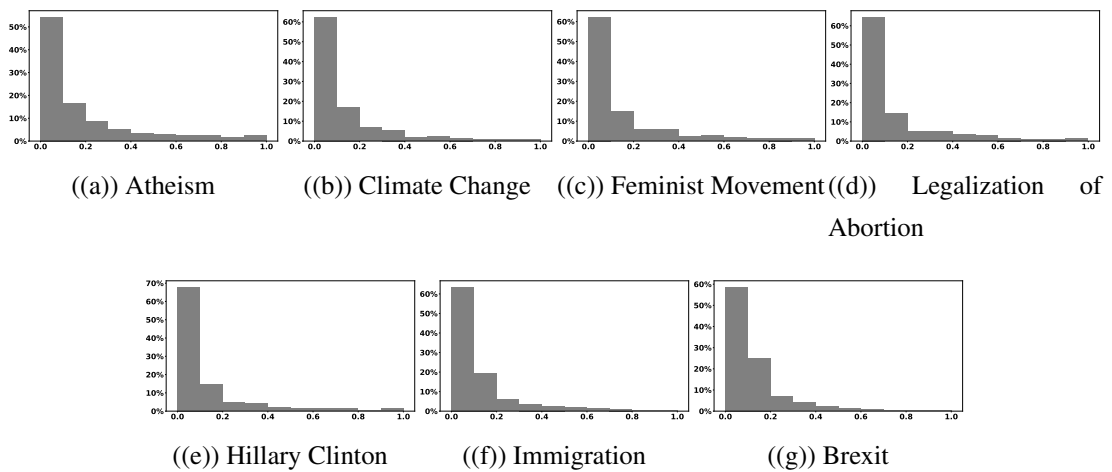


Figure C.1: The distribution of accounts scores on the top 1,000 influential accounts

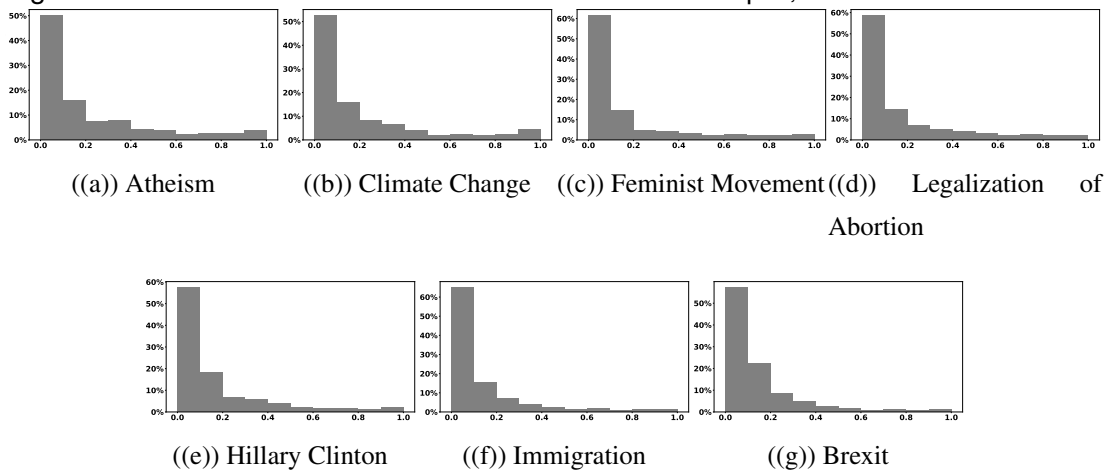


Figure C.2: The distribution of accounts scores on the top 1,000 influential accounts from in direct interactions (EXP) in predicting the Against/ Favor stance (Topic level).

C.2 Distribution of bots on topic level

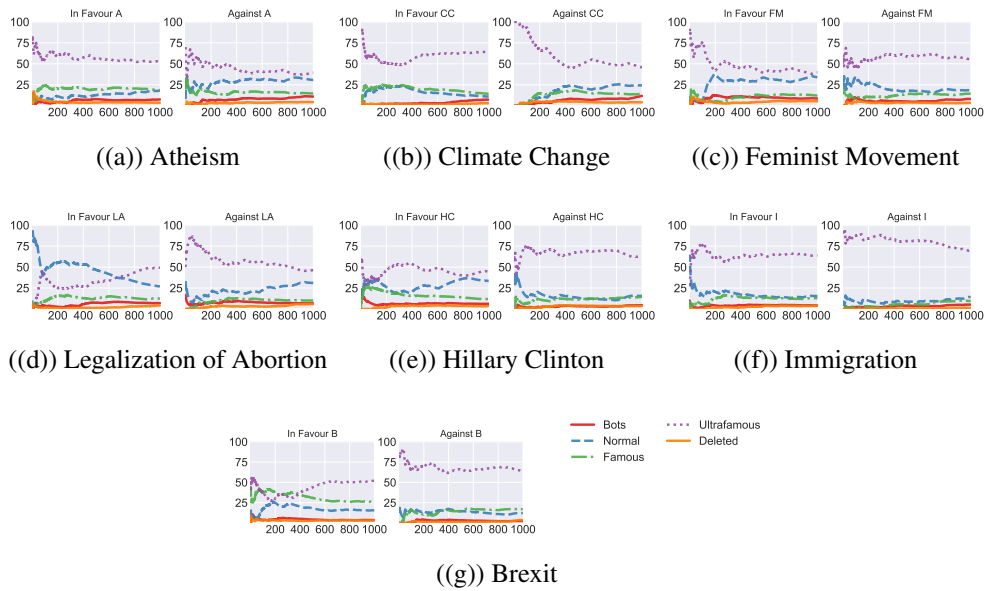


Figure C.3: The distribution of bots on the top 1,000 influential accounts from the direct interactions (IN) in predicting the Against/ Favor stance (Topic level).

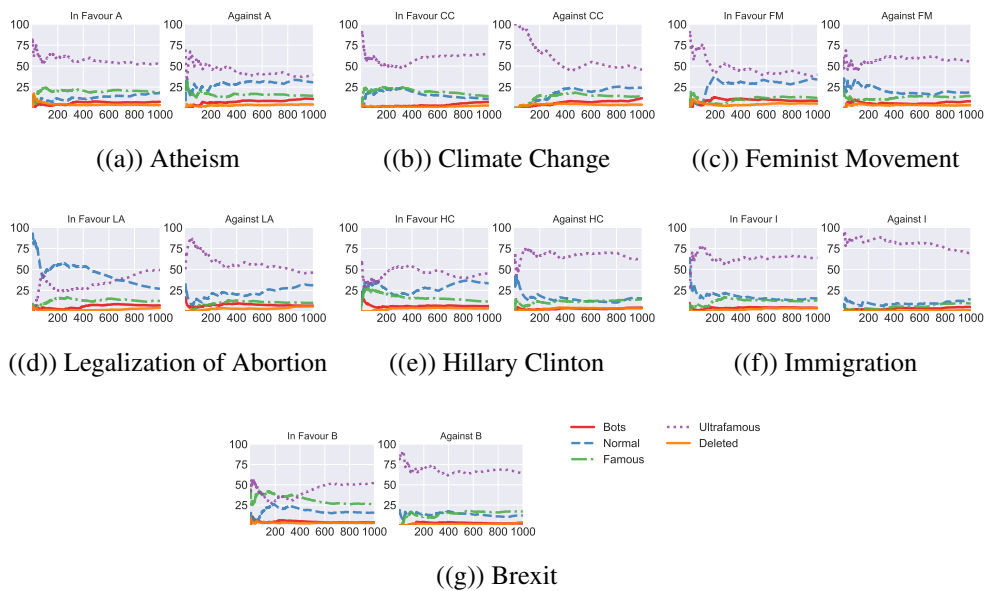


Figure C.4: The distribution of bots on the top 1,000 influential accounts from indirect exposure (EXP) in predicting the Against/ Favor stance (Topic level).

C.3 Chi-squared test for accounts distributions

The chi-squared test results for each topic.

T	Favor	Against
A	1.69e-27***	1.7e-33***
CC	1.96e-06***	1.11e-65***
HC	5.43e-34***	5.28e-44***
FM	7.60e-12***	3.03e-54***
LA	3.02e-17***	2.45e-17***
B	0.56	0.56
I	0.36	0.046*
Overall	3.17e-69***	1.77e-178***

Table C.1: Chi-squared test for accounts distributions between IN and EXP bot accounts. * $p_i < 0.05$, ** $p_i < 0.01$, *** $p_i < 0.001$