Leveraging Stances in Conversations for the Assessment of Contentious events in Twitter

Ramon Villa-Cox

CMU-ISR-22-108

July 2022

Institute for Software Research
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Kathleen M. Carley, Chair
Hirokazu Shirado
Alex Davis
Ignacio Arana
Enrique Pelaez

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Societal Computing.

The research presented in this dissertation was supported in part by the Secretaría de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT), Ecuador. Additional support was provided by the Office of Naval Research under grants No. N000141812108, No. N000141712675, and No. N000142112749. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the funding agencies.

Keywords: Stance Mining, Social Network Mining, Online Polarization, Confirmation Bias, South American Protests

"... Because if the world is round, I don't know what it means to be ahead."
F.C.

Abstract

There is currently an ongoing policy discussion regarding the impact of the observed polarization online, how it affects the spread of false information, and what if anything should be done to curtail it. To design and implement effective and efficient interventions in this area, requires a detailed understanding of how the members of polarized communities interact with each other and with outsiders holding opposing views. The focus of this dissertation is the study of polarized Twitter communities, and the spread of disinformation through them, during contentious events. Due to its large number of users, Twitter has become one of the primary social media platforms for acquiring, sharing, and spreading information. However, it has also become a source for misinformation spread and polarization. The effect might not be crucial when the subject in hand is a trivial one, however, during globally concerning events, it gains an undeniable importance. A significant amount of research on information diffusion through this medium has focused on retweeting, despite it being only one potential reaction to information found on Twitter. As shown in this work, this can be misleading, particularly when characterizing the spread of disinformation or when identifying polarized communities.

To address this, we explore two different subareas of the identification of stance in Twitter conversations, one that seeks to identify a user's stance towards a pre-defined target (target stance classification) and one that focuses on the stance to messages from other users (conversation stance classification). Analyzing such conversations is difficult and requires complex natural language processing models that often rely on copious amounts of labeled data. These issues are amplified when working in languages other than English, as labeled resources are scarcer. In the pursuit of the objectives set forth in this work, we developed a weak-labeling methodology for target stance detection which requires minimal labeling effort and constructed one of the first labeled datasets in Spanish for the identification of stance in conversations. This dataset was constructed seeking to provide a unified benchmark for the detection of both polarized online discussions and rumors. These resources are then used for the development of state-of-the-art stance classifiers to explore polarized Twitter communities during a major political event that shocked the South American Region at the end of 2019. For example, results show that a user's tendency to share information consistent to their views of the government is not consistent to the "filter bubble" explanation for polarization. That is, we show that users from both sides of the ideological spectrum actively engaged with each other (mostly negatively). This implies that the observed phenomenon is more consistent with polarized social media practices consistent with confirmation bias on part of the users.

Acknowledgments

Firstly, I would like to thank my advisor Prof. Kathleen M. Carley for her encouragement and support throughout my PhD. studies. I am also grateful to the other members of my committee for their helpful comments and suggestions, which improved the quality of this dissertation. Additionally, I would like to thank all my coauthors, namely Mathew Babcock, Sumeet Kumar, Ashique KhudaBukhsh, Helen Zeng, Evan Williams, Enrique Pelaez, and Gonzalo Villa-Cox. Much of the work presented here would not have been possible without their collaboration.

Secondly, I am grateful to the Carnegie Mellon community for providing a great environment for scientific research. In particular, I would like to thank the current and past members of the CASOS group for the helpful discussions and Sienna Watkins for her assistance throughout this process.

Finally, I would like to thank my family, starting with my parents, whose sacrifices paved the road I have traversed. My brother not only for his advice, but for always being there for brainstorming. Anahi Salazar for her companionship and the sacrifices made while I chased this dream. For this, and so much more, I will always hold you in my heart. And lastly, to my papi Ramon and tío Julio, who passed while I pursued my studies. This is for you.

Contents

1 Introduction **1**
- 1.1 Background . . . 3
- 1.2 Datasets . . . 4
  - 1.2.1 Contentious Movie Releases . . . 4
  - 1.2.2 South American Protests . . . 6
  - 1.2.3 The 2020 Chilean Plebiscite . . . 8
- 1.3 Contributions . . . 8
- 1.4 Organization of this Thesis . . . 10

2 The spread of False Information and its mitigation: The Captain Marvel and Black Panther case studies **13**
- 2.1 Introduction . . . 13
- 2.2 Related Work . . . 14
- 2.3 Data Description and Methods . . . 15
- 2.4 Results . . . 17
  - 2.4.1 Speed of diffusion by origin post, response type, and communities in the Black Panther discussion . . . 17
  - 2.4.2 How framing affects the success of boycott campaigns in the Captain Marvel discussion . . . 19
- 2.5 Limitations . . . 23
- 2.6 Discussion . . . 23

3 On Melding Network and Linguistic Polarization Methods: A Case Study on the 2019 South American Protests **25**
- 3.1 Introduction . . . 25
- 3.2 Literature Review . . . 27
- 3.3 Weak Labeling Methodology . . . 28
  - 3.3.1 Stance-Tags Validation . . . 29
  - 3.3.2 Determining User Stances . . . 30
  - 3.3.3 Ethical Considerations . . . 32
- 3.4 Linguistic Polarization . . . 32
- 3.5 Polarization in News Sharing Behavior . . . 36
  - 3.5.1 Quantifying the Polarization . . . 38
  - 3.5.2 Polarization through News Media Transitions . . . 39

3.6 Limitations . . . 41
3.7 Discussion . . . 42

4 Protest Stance Detection: Leveraging heterogeneous user interactions for extrapolation in out-of-sample country contexts 43
4.1 Introduction . . . 43
4.2 Related Work . . . 44
4.3 Data . . . 46
4.4 Methods . . . 48
4.4.1 Tweet Encoder . . . 49
4.4.2 User Encoder . . . 50
4.4.3 Network-Based Prediction . . . 50
4.5 Results and Discussion . . . 51
4.5.1 Main Results . . . 51
4.5.2 Robustness Analysis . . . 52
4.6 Discussion . . . 54
4.7 Limitations and Future Work . . . 55
4.8 Conclusions . . . 55

5 Stance in Replies and Quotes (SRQ): A New Dataset For Learning Stance in Spanish Twitter Conversations 57
5.1 Introduction . . . 57
5.2 Related Work . . . 58
5.3 Methods . . . 60
5.3.1 Dataset Collection Methodology . . . 60
5.3.2 Annotation of Sampled Conversations . . . 62
5.3.3 Proposed Classifier . . . 65
5.4 Results . . . 67
5.4.1 Annotation Results . . . 67
5.4.2 Baseline Results . . . 68
5.4.3 Consolidating the Conversation Stance Detection Task . . . 71
5.5 Limitations . . . 72
5.6 Discussion . . . 73

6 Towards a policy-oriented test-bed for the spread of Contentious Messages in Twitter 75
6.1 Introduction . . . 75
6.2 Background . . . 76
6.2.1 The Ecuadorian "Cacerolazo" . . . 76
6.2.2 Simulation of Twitter Discussions . . . 77
6.3 Methods . . . 78
6.3.1 Quantifying Observed Polarization . . . 78
6.3.2 Simulation Description . . . 79
6.4 Results . . . 83

6.4.1 Exploring the Conversation Networks . . . 83
6.4.2 Other empirical regularities . . . 85
6.4.3 Virtual Experiments . . . 87
6.5 Limitations . . . 89
6.6 Conclusions . . . 89

7 Conclusions, Limitations and Future Work **91**
7.1 Conclusions . . . 91
7.1.1 Similarities of Protest Stance and Community Detection . . . 94
7.2 Limitations . . . 95
7.2.1 Generalization of Proposed Algorithms . . . 96
7.2.2 Scalability of Proposed Algorithms . . . 97
7.3 Ethical Considerations . . . 97
7.4 Future Work . . . 98

Bibliography **101**

List of Figures

1.1 Example Twitter interaction during the 2019 Chilean protests. 2

2.1 Diffusion of retweets related to the top 5 Fake Attack origin tweets over time by community 19
2.2 Diffusion of related to the top 3 Satire Attack origin tweets over time by community 20
2.3 Diffusion of boycott/anti-Captain Marvel Alita campaigns. 21
2.4 Twitter Topic Groups Network. 22
2.5 Twitter user x Twitter user shared Youtube author network. 22

3.1 Performance of the weak labeling methodology on the labeled political figures at different probability thresholds. 30
3.2 The clusters of the news outlets. 39
3.3 Distribution of relative entropy for local and regional Russian and Venezuelan media 40

4.1 Proposed Architecture for the User Level Stance Classification. 48

5.1 Methodology used for the construction of the sample to be annotated. 60
5.2 Rendered Form for the annotation of a Quote-type response to a *contentious candidate*. The English translations are added for illustrative purposes. 63
5.3 Form presented to *Arbiters* for the resolution of Problematic Tweets. 65
5.4 Column-normalized confusion matrices for the 6-class conversation stance predictions 71
5.5 Column-normalized confusion matrices after the model averaging heuristic for the 4-class "Conversation Stance" and "Perception of Veracity" classifier. 73

6.1 Contentious Twitter conversation discussing the motivations of the Cacerolazo that took place in Quito. 77
6.2 Friend/Follower network for the users who participated in the *cacerolazo* protests. 82
6.3 Conversation networks for the *cacerolazo* that occurred on Quito on the night of the 12 of October, 2019. 83
6.4 Two views of the Response Network based on the predicted "Conversation Stance" of the interactions. 85
6.5 Average number of tweets poster by a user as a function of the number of Friends and Followers. 86

7.1 Retweet networks for the users that participated in the *cacerolazo* event. 94

List of Tables

1.1 Collection period and number of tweets collected for each country. 7
1.2 Distribution of labeled users and the count of valid Spanish tweets by their stance towards the 2020 Chilean Constitutional Referendum. 8

2.1 Community summaries 18
2.2 Number of users who switched between supporting Pro and Anti-Fake Attack tweets 18

3.1 Distribution of labeled political figures and stance hashtags. 29
3.2 Government Stance of users based on hashtag usage. 31
3.3 Government Stance of users based on endorsement of political figures. 31
3.4 Stance Distribution of users matched by both methodologies 32
3.5 Number of Weakly-Labeled Users and their original tweets (not including retweets). 32
3.6 Pairwise similarity between languages computed for: Pro and Against government communities and a baseline obtained from two random samples from the combined corpus. 34
3.7 Notable instances of linguistic polarization by topic for Ecuador and Bolivia. . . 35
3.8 Pairwise similarity between languages computed for pro-protest communities in different countries, and against-protest communities in different countries. 36
3.9 Number of news agencies in each country. 37
3.10 Distribution of the labeled tweets and resulting predictions after classification. . . 38
3.11 Transition matrix for Bolivia, Chile, Colombia and Ecuador 41
3.12 Summary Mobility Indices 41

4.1 Distribution of labeled users, their first-order neighbors and their tweets for each of the countries studied. 46
4.2 Final weak-label distribution of users based on the source used to assign the stance. 47
4.3 Distribution of labeled users and the count of valid Spanish tweets by their stance towards the 2020 Chilean Constitutional Referendum. 48
4.4 Performance of in-country Stance Classifiers at different context levels. 52
4.5 Macro F-1 (%) score for out of sample cross-country predictions for classifiers at different context levels. 53
4.6 Out of sample Predictions for the Chilean Referendum at different context levels 54

5.1 Distribution of relevant tweet pairs by response type that define the sampling universe. 61
5.2 Final Sample Distribution based on the country referenced in the tweet pair as identified after the annotation process. 62
5.3 Results for the Target Stance classification task for the Response and its target. . 68
5.4 Inter-annotator agreement at three aggregation levels measured by Cohen's κ coefficient. 68
5.5 Results for the two components of the Conversation Stance classification task. . . 69
5.6 Accuracy and Macro F1-score for the event identification, target stance classification tasks. 70
5.7 Accuracy and Macro F1-score of the *twBETO* baseline for the conversation stance classification tasks. 71

6.1 Polarization metrics for different user interaction networks during the event. . . . 84
6.2 Polarization metrics for the different types of user response networks based on the predicted conversation and rumor stance of the edges. 85
6.3 Virtual Experiment table for a 4-3-2 experimental design. 87
6.4 Random Walk Controversy score for the different simulated user interaction networks. 88
6.5 Summary of Stylized Facts 88

7.1 Computational cost (in hours) of the different classifiers proposed based on the GPU used. 97

Chapter 1

Introduction

There is currently an ongoing policy discussion regarding the impact of the observed polarization online, how it affects the spread of false information, and what if anything should be done to curtail it. Options have been presented in academia and the public discourse ranging from those that focus on government and platform-based interventions to those seeking to educate and empower individuals in their interactions with social media [79]. Others have investigated community-based options for fact checking or moderation and reporting bad actors [7]. To design and implement effective and efficient interventions requires a more detailed understanding of how the members of polarized communities interact both with each other and with outsiders holding opposing views. The differences in their behavior can further affect how false information and the response to it spreads through these types of communities. This understanding is complicated by the reality that there are different kinds of false information being shared from and through different online communities [116]. Features important for intervention-related decisions, such as the overall reach and speed at which false information diffuses may be dependent on the type of story and the source and target communities. It may also be affected by the diffusion of negative responses to the false information that occur prior to the news about these claims going more public [11].

Due to its large number of users, Twitter has become one of the primary social media platforms for acquiring, sharing, and spreading information. Nonetheless, it has also inevitably become a source for misinformation spread and controversy, which immensely affects what masses believe to be the truth [126] and how they behave on a day-to-day basis. The effect might not be crucial when the subject in hand is a trivial one albeit, during globally concerning events, the spread of information through Twitter gains an undeniable importance. A significant amount of the research on information diffusion in social media has focused on retweeting. However, there are many reasons why people retweet [89] and retweeting is only one potential reaction to information found on Twitter. Examining simple retweet totals is static, and doing so does not provide insight into the entire diffusion path [5]. Identifying the full temporal path of information diffusion in social media is complex but necessary for a fuller understanding [106]. Moreover, people often use other mechanisms such as quoting to resend messages [20]. In doing so, they can change the context of the original tweet, as when quotes are used to call out and attack the bad behavior of the original tweeter [10]. Examining replies to tweets and the support of those replies also assists in creating a more complete picture of the discourse path. As shown

Figure 1.1: In their interactions Twitter users often reveal their stance not only towards their target, but also to specific events or actors. In this example taken during the 2019 Chilean protests, a congressperson denounces an instance of police abuse during the protests while a another user accuses him misconstruing the events. The replying user not only reveals his opposition to the protests taking place, but this stance also informs her negation of the events described in the target tweet.

in this thesis, focusing solely on retweets or aggregating any type of response to a tweet can be misleading, particularly when characterizing the spread of disinformation or when identifying polarized communities. We find that negative responses towards fake tweets tend to manifest through replies and quotes and that diffusion through this medium can dwarf what is observed through retweets. For this reason, its important to exploit community responses to false information and rumors in order to improve their early detection.

As a matter of fact, users can convey supporting or opposing stances when replying or quoting a tweet. Inferring stances in such responses, as in whether the response post favors or opposes the original post, is increasingly being used not only to detect online harassment [83] but also for identifying misinformation [141], and therefore, is an important research direction. Still, the majority of resources for this task are built for rumor detection and do not generalize to non-rumor events [22] which limits their usefulness for assessing polarization or controversy. As an illustrative example, Figure 1.1 shows a Reply interaction between users during the protests that paralyzed Chile in 2019. Here a user questions the existence of police abuse reported by a sitting congressperson (the source tweet) by arguing that the video omitted the events that led to the officer's behavior. In doing so, the user not only reveals her opinion of the events that transpired but also her opposition towards the protests that were taking place. Analyzing these type of responses may allow a better estimate of the perceived public support for the protests and government during charged political events. Importantly, as we show in this thesis, we can also leverage the stance revealed by users during the 2019 Chilean protests to successfully predict their support for the Constitutional Plebiscite that took place at the end of 2020.

This dissertation explores two different subareas of the identification of stance in Twitter conversations, one that seeks to identify a user's stance towards a pre-defined target (*target stance classification*) and one that focuses on the stance to messages from other users (*conversation stance classification*). Analyzing such conversations is difficult, and requires complex natural

language processing (NLP) models that often rely on copious amounts of labeled data through supervised learning. These issues are amplified when working in languages other than English, as labeled resources are scarcer. In the pursuit of the objectives set forth in this work, we developed a weak labeling methodology for target stance detection in Twitter data that requires minimal labeling effort, and constructed one of the first labeled datasets in Spanish for the identification of stance in conversations. Moreover, this dataset was constructed seeking to provide a unified benchmark for the detection of both polarized online discussions and rumors. These resources are then used for the development of state of the art stance classifiers to explore polarized Twitter communities, and the spread of rumors through them, during contentious events. In particular, we focus on a major political event that shocked the South American Region at the end of 2019, and whose consequences still shape the current political landscape [130]. Results show, for example, that a user's tendency to share information consistent to their views of the government is not consistent to the "filter bubble" explanation for polarization. That is, we show that users from both sides of the ideological spectrum actively engaged with each other (mostly negatively) during a particular polarizing event. This implies that the observed phenomenon is more consistent with polarized social media practices consistent with confirmation bias on part of the users.

1.1 Background

Stance detection is an essential component of many tasks associated with online social network moderation and analysis, including propaganda, misinformation, hate-speech detection and even opinion polling. In some domains, sentiment analysis may be a reasonable approximation of stance, but it has been shown that sentiment polarity is a poor proxy, particularly around online discussions of contentious political issues [92]. Du Bois defines the act of stance taking as ''*... a public act by a social actor, achieved dialogically through overt communicative means, of simultaneously evaluating objects, positioning subjects (self and others), and aligning with other subjects...*'' (page 163, [37]). Work in this area has concentrated in exploring stance in conversations and on debates with respect to a predefined topic or target (known as target-stance classification). However, progress in this area has been limited by its reliance on small hand-labeled datasets, primarily created around challenge competitions like SemEval-2016 [92]. Getting labeled data is expensive as conversations can be about any topic, topics change over time, and new topics emerge, all of which makes learning the stance from conversations a challenging problem. Researchers have already created labeled datasets for a few controversial topics [61, 91]. However, labeled examples are scarce and are only available for a few topics, so training complex models on newer topics is still difficult. Importantly, even less resources on this area are available in languages other than English and, to the best of our knowledge, for this thesis we construct the first hand-labeled dataset for stance in Spanish Twitter conversations. In what remains of this section we provide a brief overview of the main research that have been applied in this area. For a more detailed description of these areas, and main datasets available, we refer the reader to Küçük and Can [75].

Target Stance Classification Target Stance Classification focuses on classifying the stance of a user or document with respect to a predefined topic or target. Task 6 of Semeval 2016 is also a common benchmark for target stance classification. Authors have achieved SOTA results on this benchmark using architectures ranging from end-to-end neural ensemble models [108] to hand-crafted feature-based classifiers [3]. Due to the limited amount of data available, algorithms which rely on hand-crafted features are still prominent and achieve competitive performances with Deep Learning algorithms.

Stance in Conversations Conversation-Stance classification focuses on identifying the stance of responses in a conversation using "deny", "support", "comment", or "query" labels. Most work in this area has focused on leveraging these interactions for the identification of rumors (also known as rumor-stance classification) [88, 141, 143]. This focus has in large part been driven by the resources made available for this task in challenge competitions like PHEME [72] and the 2016 SemEval [91]. Though useful for rumor detection, this does not generalize to non-rumor events [22] and limits their usefulness for assessing polarization. Despite this, a largely independent research vein has also recognized the utility of leveraging stance for the detection of controversy [2, 75].

A note on terminology The different studies tackled in this thesis are centered around contentious events and polarized discussions. I adopt a common definition for this phenomena which is centered in controversy. Controversial events tend to engage large social media audiences and elicit public discussion were audience members express opposing views or disbelief [99].

1.2 Datasets

In this section we present an overview of the datasets constructed and used throughout this thesis.

1.2.1 Contentious Movie Releases

Black Panther We use Twitter data related to the opening weekend (15–18 February 2018) of the Marvel superhero movie Black Panther. Black Panther was a financial and critical success that was promoted in part by its status as the first Marvel Cinematic Universe movie to have a predominately African and African-American cast and focus. The opening of the Black Panther movie makes for a good case study because of the high level of Twitter activity surrounding the movie (is was reported as the most tweeted about movie Twitter 2018) and due to the presence of multiple types of disinformation campaigns within the Twitter conversation.
The data set contains approximately 5.2 million tweets related to Black Panther which were collected from 8 February to 16 March 2018 using Twitters public API. In previous work [10], four types of false information stories were identified: (1) Fake Attack posts claiming racially-motivated physical violence at movie theaters which were debunked, (2) Satirical Attack posts making similar but more exaggerated claims in an apparent attempt to mock or shame the original Fake Attack posts, (3) Fake Scene posts claiming the film contained scenes (mostly racially-inciting), that it did not and (4) Alt-Right posts claiming the movie was supportive of Alt-Right

ideology (in the film such policies are questioned and repudiated). The authors identified a total of 304 origin posts (tweets with an original false claim of one of the 4 types) and approximately 155,000 tweets that responded (retweeted, quoted, or replied) to those origin posts.
Accepting retweets as endorsements, we manually verified which replies and quotes where endorsing and which ones were detracting from the origin post being replied to/quoted. Though some origin posts garnered hundreds to thousands of replies or quotes, only 89 quotes and 17 replies garnered more than 10 retweets, and we restricted our verification to those responses. This dataset is used primarily in **Chapter 2**.

Captain Marvel Marvel Studio's first female lead-focused superhero movie, was released on March 8, 2019. Much of the Twitter discussion of the movie was standard comic book movie corporate and fan material. However, in the runup to the release of the movie there was also a significant amount of contentious discussion centered on whether the movie should be supported and/or whether illegitimate efforts were being made to support or attack it. From February 15 to March 15, 2019, we used Twitter's API to collect tweets for our analysis. Our goals and methods for tweet collection were as follows:

1. Compare the origination, spread, and response to the two main campaigns to push the misinformation campaign on interest. To do this we collected all non-reply/non-retweet origin tweets that used #BoycottCaptainMarvel and #AlitaChallenge during our period of interest and collect all quotes, replies and retweets of these origins.
2. Explore the conversation around these two campaigns that did not use the two main hashtags. Compare the hashtag-based and non-hashtag-based conversation. To do this we collected all non-reply/non-retweet origin tweets that used "Alita" along with one of a set of keywords used in the contentious comic-book Twitter discussions (e.g. "SJW", "Feminazi") and collect all quotes, replies and retweets of these origins. We labeled this as the "Charged Alita" conversation.
3. Characterize the communities that users who spread or responded to the misinformation were from. To do this we collected the Twitter timelines of the central users and all non-reply/non-retweet origin tweets that provide information about the general Captain Marvel movie conversation (using the keywords #CaptainMarvel, Captain Marvel, Brie Larson), and all quotes, replies and retweets of these origins. To use cross-platform information to understand community structure we collected author, subject, and viewership information for all YouTube video URLs shared through Twitter.

For goals 1 and 3 we are relatively confident that our method allowed us to collect the vast majority, if not the entirety of the tweets, as we aimed to focus on very specific hashtags and obtain a general sense of where such hashtags were used compared to the most general conversation. For goal 2, while we are able to collect a large enough sample of tweets related to the campaign, it is probable that some discussions of the Alita Challenge took place on Twitter using keywords we did not search for, and therefore are not part of our analysis. We rehydrated any available target of a reply that was not originally captured in the first collection. This allowed us to capture at least the first level interaction within the relevant conversations. In total, we collected approximately 11 million tweets. This dataset is used primarily in **Chapter 2**.

1.2.2 South American Protests

Background In 2019, a series of protests shocked the region of South America. They started in Ecuador with Chile, Bolivia and Colombia soon following. With the exception of Bolivia, the protests resulted from left-wing movements seeking to resist austerity measures being imposed in each country, or the rising costs of public services. In Bolivia, they were a right-wing response to an alleged electoral fraud undertaken by the government in favor of the president who was seeking reelection. The protests also had in common a massive online presence and the reported involvement of international and regional actors that sought to influence their evolution. These include international news agencies like RT en Español, funded in part by the Russian government, or TeleSUR and NTN24, funded in part by the Venezuelan government, that were more critical of local governments (except for Bolivia) and provided more favorable coverage of the protesters. In contrast, local news agencies tended to be more critical of them and favorable towards the government[1]. To better contextualize our work, we first present a brief overview of the main events that transpired in each of the countries.

Ecuador Protests started in October 3, 2019 as a response to an austerity package (the 883 Decree) which involved the removal of fuel subsidies. Protests leaders included indigenous movements (CONAIE) and followers of former president Rafael Correa. After two weeks of violent clashes, the temporary reallocation of the seat of government, and the paralyzation of large part of the economy; President Moreno agreed with indigenous leaders to withdraw the 883 Decree. We used 191 terms and hashtags for the collection, that included for example: #EcuadorEnCrisis, #ParoNacionalYa, #EcuadorEnResistencia, #ToquedeCacerolazo, etc.

Chile Chile is one of the wealthiest countries of South America but also one with highest inequality. On October 7, 2019 protests started because of a rise in the cost of subway tickets in Santiago (the capital), which lead to clashes between the police and protesters. Overtime this translated to a demand for structural change, and for constitutional reform. On November 15, the National Congress signed an agreement to hold a national referendum to rewrite the constitution. The protests continued well into 2020, and due to COVID the referendum was rescheduled to October 2020, when it was overwhelmingly approved with 78% of the vote. We used 244 terms and hashtags for the collection, that included for example: #ChileEnHuelga, #ChileProtests, #YoNoMarcho, #ToqueDeQuedaYA, #FueraPiñeraDictador, #ChileDesperto, etc.

[1]Lara Jakes, "As Protests in South America Surged, So Did Russian Trolls on Twitter, U.S. Finds" New York Times, January 19 2020, accessed December 19 2020, https://www.nytimes.com/2020/01/19/us/politics/south-america-russian-twitter.html

Bolivia
: Former President Evo Morales, who was the longest-serving leader in South America with 13 years in office, was accused of wrongdoing in his fourth term election. On October 21, protests started around his reelection and demanding the nullification of the elections. On November 10 an audit team from the Organization of American States (OAS) questioned the integrity of the election. The same day following pressure from the armed forces, Evo Morales announced his resignation. Protests continued until the end of November, primarily by those that sought Morales' return. We used 244 terms and hashtags for the collection, that included for example: #EleccionesBolivia2019, #BoliviaDiceNo, #GolpeDeEstadoEnBolivia, #FraudeElectoralEnBolivia, #EvoEsDemocracia, etc.

Colombia
: The protests in Colombia were a response to economic and political reforms proposed by President Ivan Duque. On November 21, massive protests started throughout the country demanding the end of austerity measures. Protesters displayed flags of Chile and Ecuador, banners reading "South America woke up", and chanted anti-violence slogans. The protests continued throughout the year, with multiple clashes between the public and the armed forces. We used 201 terms and hashtags for the collection, that included for example: #CarcelParaPetro, #ApoyoALaFuerzaPublica, #ColombiaDesperto, #ESMAD, #ParoNacional25N, etc.

Data collection The dataset consists of 100 million tweets from 15+ million users collected using Twitter's API v1 around the protests that transpired in Ecuador, Chile, Bolivia and Colombia. For each event, we built the queries by first identifying at least 12 of the most prominent hashtags/terms (using Twitter's trending terms in the country). After some days of streaming, we determined the most frequent hashtags not yet included that were relevant to the protests, taking special effort to include hashtags that were used by different groups (for and against the different governments). We included these to our query and every 7 days we performed REST grabs of all the terms (to ensure their collection from the start). By repeating this process each week, we built up the set of more than 500 hashtags. To improve the quality of the conversational structure present in the data, we also re-hydrated any missing targets or ancestors (up to 5 levels above in the conversation tree) of replies or quotes. Table 1.1 presents other descriptive statistics from the data collected for the different countries.

This dataset serves as the basis for the construction of the weak-labeled dataset described in **Chapter 3**, which in turn is used in **Chapter 4**, and for the construction of the hand-labeled dataset covered in **Chapter 5**. Both of these resources are used in **Chapter 6**.

	Collection Period (in 2019)	**Number of Tweets (Millions)**
Bolivia	October 15 to November 24	23.5
Chile	October 10 to November 24	59.6
Colombia	November 10 to December 24	20.4
Ecuador	September 25 to October 24	19.1

Table 1.1: Collection period and number of tweets collected for each country.

1.2.3 The 2020 Chilean Plebiscite

Throughout the 2019 Chilean protests, different social movements made calls in favor of drafting a new constitution. This social pressure reached a boiling point in November of that same year, which led the Chilean national congress to agree to hold a National Plebiscite. It was overwhelmingly approved with 78% of the vote in October of 2020. Using Twitter's v2 full-archive search endpoint feature available on their Academic Research Track[2], we collected tweets from September 25 to November 10 of 2020 (a month prior and two weeks after the plebiscite took place). We collected tweets matching 124 hashtags and terms relevant to the event and, following the methodology established in **Chapter 3**, assigned weak labels denoting whether a user supported or opposed drafting a new Constitution. Table 1.2 provides descriptive statistics of the labeled referendum users. We note that 45.5% of users labeled are not included in the 2019 Chilean Protest data. Moreover, to improve the resolution of the networks available for the labeled users, we collected their timeline during the event. Timelines were not always collected in the protest data, so we hypothesize that the additional timeline context for each user will improve the quality of user embeddings for the referendum data. This dataset is used in **Chapter 4** and further details of the labeling methodology are presented there.

	Against	**Pro**	**Neighbors**
Users	10,423	11,206	96,202
Tweets	6,454,647	6,060,679	1,288,351

Table 1.2: Distribution of labeled users and the count of valid Spanish tweets (including retweets) by their stance towards the 2020 Chilean Constitutional Referendum. We also include counts corresponding to their first-order neighbors (based on the response network).

1.3 Contributions

The contributions of this thesis can be summarized as follows.

- ***Methods*:** First, we exploit community interaction (via Replies and Quotes) to explore the response of different communities to disinformation (Chapter 2). We find that negative responses towards fake tweets tend to manifest through replies and quotes and can significantly affect the diffusion of a fake story. This is important considering that most research has characterized the diffusion of rumors by focusing on retweets or by aggregating all responses to them which can be misleading when characterizing the spread of disinformation.
  We also develop a novel method to mine stance in Twitter conversations that requires minimal supervision and leverages users' endorsement of politicians' tweets and hashtag campaigns with defined stances towards the protest (for or against) (Chapter 3). The reliance

[2]This allows full historical access to publicly available tweets matching complex queries.

on not only hashtags, but also endorsement of political figures provides different avenues to assess the robustness of the labels obtained and ameliorate the problem of hashtag hijacking. The mined stances are used to segregate the user pool into two groups: one in favor of the government, and the other, against it. The method relies on a weakly labelled approach and thus do not require a large number of labels. It is validated by showing that the constructed labels partition the users in communities that are polarized in their language and news sharing behavior. To the best of our knowledge, this work is one of the first few that combines network-based methods and language-based methods to jointly explore the nature of polarization in a user's news sharing behavior and her language usage. Moreover, the methodology holds promise for the development of large-scale databases for the analysis of similar contentious events (with the active involvement of local political figures).

Finally, we propose a unified approach for the annotation of stance in conversations (Chapter 5). Current approaches are tailored for rumor detection which limits their usefulness for assessing polarization or controversy [22]. For this reason, we separate the task of detecting stance in conversations in two parts: the first which seeks to identify (dis)agreement or neutral responses, and the second which identifies whether the agreement (disagreement) is actually supporting (denying) the veracity of said statement. We believe that this unified treatment can help bridge these two related research areas and is necessary in order to explore how polarization in online discussions can affect the spread of rumors.

- ***Resources*:** We release a dataset of over 15 million tweet IDs with the weakly labeled stance of approximately 500k users collected around the 2019 South American Protests (Chapter 3). To the best of our knowledge, no large scale social media dataset relevant to protests spanning multiple countries exist. Similarly, we also release a dataset of 10 million tweet IDs with 20k weakly labeled users collected around the 2020 Chilean Plebiscite (Chapter 4). Finally, we hand-labeled a sample of conversations from the collection of the 2019 protests and release the first Spanish dataset for the classification of the stance in conversations (Chapter 5).
- ***Algorithms*:** We propose a large-scale stance-detection architecture using transformers and graph neural networks that leverages a user's social network, a user's tweet timelines, and the weakly-labeled protest-related tweets in order to predict their stance towards the government of each country (Chapter 4). By exploiting the regional nature of the dataset, we are able to show how increasing the context available for an algorithm not only improves its local performance but also greatly enhances its ability to extrapolate to new country-contexts and to future data. This is salient, as exploring the effect of context on the performance of algorithms has been identified as important open problem in the area of stance classification [75].

  In addition, we develop a variant of a BERT [32] language model specialized on Spanish Twitter conversations and train it on a substantial corpus of approximately 200 million tweets (Chapter 5). This is used to predict the conversation-stance of users in the labeled dataset, and we are able to achieve a better performance than other Transformer-based classifiers. Finally, we develop an agent-based dynamic-network model validated by leveraging the predictions of the aforementioned classifiers.

- ***Social*:** We look at a globally important event: the series of South American protests that took place in multiple countries in 2019. We observe that linguistic polarization mainly manifested along ideological, political or protest-related lines. Moreover, we find strong evidence of the polarization in users' news sharing patterns, consistent with their stances towards the government. These can have pervasive effects on public discourse and political literacy. In this vein, we show that these practices are not consistent with the existence of "filter bubbles" as users actively engaged with others (mostly negatively) holding opposing views. Finally, we show that the stance revealed by users during the protests can be predictive of their position during future electoral processes (as we show in the case of the Chilean Plebiscite).

1.4 Organization of this Thesis

This document is organized as follows:

1. Introduction
2. The spread of False Information and its mitigation.
3. On Melding Network and Linguistic Polarization Methods.
4. Leveraging heterogeneous user interactions for extrapolation in out-of-sample country contexts.
5. A New Dataset For Learning Stance in Spanish Twitter Conversations.
6. Towards a policy-oriented test-bed for the spread of Contentious Messages in Twitter.
7. Conclusions and future work.

The thesis can also be divided in 3 parts. The first part, comprised of chapters one through three, serves as motivation for the work undertaken, and presents case studies exploring empirical regularities of disinformation campaigns and polarization on two different types of contentious events. **Chapter 1** (this chapter) introduces the work and provides the background and contributions of this thesis. In addition, it also describes the datasets used in the subsequent chapters. **Chapter 2** presents a compilation of two case studies that explored different aspects of disinformation campaigns that targeted the release of two popular Marvel movies, first the diffusion of false stories and their responses, and second the effect of framing on the success of the campaigns. We find that the negative reaction to fake stories of racially-motivated violence, whether in the form of debunking quotes or satirical posts, can spread at speeds that are magnitudes higher than the original fake stories. Overall, this work helps to illustrate the importance of investigating "on-the-ground" community responses to fake news and other types of digital false information and to inform identification and intervention design and implementation. In **Chapter 3** we analyze an important sociopolitical event, the 2019 South American protests, and explore online polarization focusing on two dimensions: Polarization in language and in news consumption patterns. We demonstrate that (1) a combined treatment offers a more comprehensive understanding of the event; and (2) these cross-cutting methods can be applied in a synergistic way.

The insights gained by the combination of these methods include that polarization in users' news sharing patterns was consistent with their stances towards the government and that polarization in their language mainly manifested along ideological, political or protest-related lines. In addition, we release a massive data set of 15 million tweet IDs relevant to this crisis with the weakly labeled stance of approximately 500k users. Importantly, this two chapters emphasize the roll that interactions other than retweets, like quotes and replies, can have on assessing polarization or predicting the diffusion of disinformation on Twitter.

The second part, comprised of chapters four and five, is focused on developing neural algorithms for the tasks of target-stance and conversation-stance classification. **Chapter 4**, explores the performance and extrapolation power of political stance-detection models (a target-stance detection task) using the weakly-labeled stance dataset constructed in the previous chapter. We develop transformer-based user and tweet encoders to embed users in a low-dimensional spaces using their text and social networks. We then train heterogeneous graph attention networks to predict user stances, and contrast the ability of the network and transformer-only models to extrapolate stance predictions to different country-contexts. We find that leveraging users' social networks in stance detection improves in-country model performance for every country examined. More notably, we find that our heterogeneous graph neural network models, which differentiates between retweet and response links, greatly enhance the ability of the stance detection models to extrapolate to new country-contexts and to future data. The latter is done by utilizing the classifiers trained on the Chilean protest data to predict the results of the Plebiscite that occurred the following year. In **Chapter 5**, we create the first conversation-stance dataset in Spanish, by labeling tweet pairs collected during the South American protests. We separate the task of detecting stance in conversations in two parts: first we identify whether the response agrees, disagrees, comments or queries its target, and then we determine whether the agreement (disagreement) is actually supporting (denying) the veracity of said statement. In addition, we also identify the stance of the target and response tweets towards the protests and government in each country (a target-stance detection task). We then develop a variant of a BERT language model specialized on Spanish Twitter conversational data and train classifiers of the conversational-stance of users and and their perception of rumours that diffused throughout the event.

The final part, comprised of chapters six and seven, serves as a coda to this dissertation. **Chapter 6** presents a case study focused on the "cacerolazo" (pot and pan banging) protests that took place during government-mandated curfews in the height of the 2019 Ecuadorian protests. This event is specially relevant to the themes explored in this thesis, as people both supporting and opposing the government partook in it while confined to their homes and actively claimed in Twitter that it supported their position. We leverage the classifiers developed in the previous chapters to identify the protest-stance of users active during the discussions and also the stance of their interactions. We find that even though the Retweet user network is highly polarized, others like the Friend or Response networks, show little or no polarization. This implies that users from both sides of the ideological spectrum actively engaged with each other during the event. This suggests that the observed polarization is not consistent with filter bubbles and is more reflective of confirmation bias on part of the users. These empirical regularities are then used to validate

an agent-based dynamic-network model based on the event. Via a virtual experiments, we show that the topology of the empirical Friend/Follower networks are not sufficient to recreate the observed levels of polarization, while our proposed mechanisms for confirmation bias are able to do so. This is meant to serve as a proof of concept of how these methods can be combined in a synergistic way to develop policy-oriented test-beds, grounded on real data, for these social media sites. Finally, **Chapter 7** concludes, and I describe the limitations of this work and suggest directions for future research.

Chapter 2

The spread of False Information and its mitigation: The Captain Marvel and Black Panther case studies

2.1 Introduction

Inaccurate and misleading information on social media can be spread by a variety of actors and for a variety of purposes. In potentially more immediately higher stakes Twitter conversations such as those focused on natural disasters or national elections, it is taken for granted that pressure groups, news agencies, and other larger and well-coordinated organizations or teams are part of the conversation and may spread misinformation. In less directly political or emergency-related conversations, there remain questions of how such organized misinformation campaigns occur and how they shape discourse. For example, the Marvel Cinematic Universe comic book movies are very popular and the most financially successful film franchise to date. As the Marvel movies have expanded their casts and focused on more diverse storytelling and storytellers, they have become flash-points in the United States for online expressions of underlying cultural debates (both those debated in good faith and those that are not). This trend appears to be part of a wider one in the comic book communities on social media and within comics' news and politics sites [102].

Social media platforms can be places where honest cultural and political debates occur. However, because of the presence of misinformation and propaganda, many such conversations are derailed, warped, or otherwise "polluted". Conversations that tend to polarize the community often carry misinformation [103], and such polluted content serves the purpose of the influencer or the creator of misinformation. Therefore, further investigation into the use of misinformation on Twitter, the way different types spread and the groups that push it in the service of cultural debates may be useful in informing community, individual, and government responses for preserving healthy discussion and debate online. To inform future work in this area, we conducted two different case studies to explore a set of these issues as they appear in the Black Panther and the Captain Marvel movies Twitter discussions, and in this chapter we present the main results obtained on each of them.

The first study [11] focused on *Black Panther*, which was a financial and critical success that was promoted in part by its status as the first Marvel Cinematic Universe movie to have a predominately African and African-American cast and focus. The opening of the Black Panther movie makes for a good case study because of the high level of Twitter activity surrounding the movie (is was reported as the most tweeted about movie Twitter 2018) and due to the presence of multiple types of disinformation campaigns within the Twitter conversation.
The second study [12] concerned *Captain Marvel*, Marvel Studio's first female lead-focused superhero movie, was released on March 8, 2019. The underlying core of the contentious discussions around the movie was related to recurring debates over diversity and inclusion in mass media in general and in comics in particular [102]. There were several different types of false or misleading information that was shared as part of these discussions. Fake claims about the lead actress and Marvel were promoted in multiple ways and were the basis for campaigns calling for a boycott of the movie, either directly (#BoycottCaptainMarvel) or indirectly (by promoting an alternative female-led action movie, Alita: Battle Angel through a hijacked hashtag, #AlitaChallenge. As these campaigns were found to have somewhat similar community foundations but experienced differing levels of success, they make for a good case study to explore the impact of organization and narrative framing on the spread of misinformation.
In this work, we seek to explore two different aspects of disinformation campaigns, encompassed by each of the case studies. In the Twitter conversation regarding the release of the Black Panther movie, we explore the diffusion of different false information stories and responses to them that occurred during the same event. In the captain marvel case study however, the sources of misinformation were more well-known organized groups and/or more established communities providing an additional avenue of exploration. Our research goals are to (1) identify communities through which the stories and responses diffused, (2) compare how fast the different types of stories and responses diffused overall and through each community, (3) compare the different origins and actors involved in two misinformation-fueled boycott campaigns, and (4) how does the framing of a disinformation campaign affect its successful diffusion.

2.2 Related Work

The rate of diffusion of information in networks depends in part on the network topology. Some network characteristics are known to help faster information diffusion whereas others inhibit the flow of information [63]. Many information diffusion studies in social networks have borrowed ideas from the models for the spread of infections [70, 135, 138, 139]. In the context of the spread of disinformation, Vosoughi et al. [126] contrasted the spread of news stories on Twitter which had been fact-checked as true and false by third parties, finding that false articles (particularly as they relate to political events) spread farther than their counterparts.
Some researchers have suggested that the differences in the way true and false information propagate could be used to detect false information [27]. However, most of the studies in this topic have characterized the diffusion of tweets by either focusing solely on retweets or by aggregating any type of response to them. As we show in the present work, this can be misleading, particularly when characterizing the spread of disinformation. We find that negative responses towards fake tweets tend to manifest through replies and quotes and that diffusion through this medium

can dwarf what is observed though retweets. Moreover, the effect of these negative responses can have an important effect on the spread of other fake tweets in the same topic, which is necessary to consider when predicting the spread of later instances of a fake story. Additionally, not all false information spread online is the same. Digital social media is the target and source for variations ranging from satire and parody to organized disinformation campaigns [116]. These types of false information differ in their purposes, design, and impact. They also can interact, such as when satire is used to mock misleading political arguments.

Lumping different types of false information together into a false/true dichotomy as much previous research has done may miss differences that are important to why and how a particular type succeeds or fails. Another important part of understanding the flow in information in networks is to understand the types of communities in a network. Communities in a network can refer to groups of nodes that are more strongly connected compared to the rest of the nodes. While network-based community detection is important in determining which types of communities are more or less susceptible to the diffusion of false information, comparison of how false information flows through known communities can also be helpful. Though there is a lot of prior research on information diffusion and network characteristics, not many researchers have explored the effect of different types of communities on the flow of information and false information. Vicario et al. [124] tried to connect misinformation spread and polarized communities. Using a framework that identifies polarizing content, they can predict fake-news topics with 91% accuracy.

Many characteristics of message content may impact information diffusion [41]. For example, moral-emotional language leads messages to diffuse more within groups of like-minded individuals [21]. Emotionally charged messages are more likely to be retweeted [113] as are those that use hashtags and URLs [114]. The use of URLs shows the cross-platform nature of misinformation on social media with many URLs pointing to other sites such as YouTube. Research only focusing on one ecosystem may mischaracterize both the who and how of misinformation spread. While our focus is on Twitter activity in the present work, we also use Twitter-Youtube connections to better understand the communities involved.

2.3 Data Description and Methods

For this study, we collected Twitter data related to the opening weekend of the two Marvel Blockbusters described before, Black Panther and Captain Marvel. There are several ways within Twitters system in which a user can respond to a tweet they are exposed to. The user can retweet (copy the origin tweet with no commentary, which is assumed to be an endorsement), reply (write new content about the origin tweet that can be endorsing, neutral or detracting from the origin tweet), or quote (copy the origin tweet with added commentary that can be endorsing, neutral or detracting). The replies and quotes in turn can be replied to, quoted, or retweeted. A user can also follow the origin tweet poster and/or like the origin post (the exact date and time of a follow or like action are not accessible using the public Twitter API, and therefore these actions are not investigated in this work). We provide a brief description of the main methods used for its analysis and the dataset collected for each event (for a more complete description refer to Section 1.2).

Black Panther The data set contains approximately 5.2 million tweets related to Black Panther which were collected from 8 February to 16 March 2018 using Twitters public API. In previous work [10], four types of false information stories were identified, but we focus on the following three: (1) Fake Attack posts claiming racially-motivated physical violence at movie theaters which were debunked, (2) Satirical Attack posts making similar but more exaggerated claims in an apparent attempt to mock or shame the original Fake Attack posts, (3) Fake Scene posts claiming the film contained scenes (mostly racially-inciting), that it did not. For the analysis presented in this paper, we choose to focus on the origin posts that had at least 50 retweets. The reason for this is that we want to have enough data points to compare the speed of the diffusion (# of retweets over time). There were 11 origin posts that met this criterion: 5 of the Fake Attack type, 3 of the Satire Attack type, and 3 of Fake Scene type. Accepting retweets as endorsements, we manually verified which replies and quotes where endorsing and which ones were detracting from the origin post being replied to/quoted. Though some origin posts garnered hundreds to thousands of replies or quotes, only 89 quotes and 17 replies garnered more than 10 retweets, and we restricted our verification to those responses. Due in part to the tweet collection methods used for this data set (keyword-based search with limited timeline data), the user networks that could be derived from them were too sparse for network-based community detection to work effectively. We therefore separated the users in our data set into three communities based on their Twitter activity related to the false information origin posts: Pro-Fake, Anti-Fake, and Mixed. A Pro-Fake user was defined as a user that exclusively does the following: tweets a Fake Attack or Fake Scene origin post, tweets a quote or reply that supports a Fake Attack or Fake Scene origin post, or retweets any of those origin posts, quotes, or replies. An Anti-Fake user was defined as a user that exclusively does the following: tweets a Satire Attack origin post, tweets a quote or reply that supports a Satire Attack post, tweets a quote or reply that attacks a Fake Attack or Fake Scene origin post, or retweets any of those origin posts, quotes, or replies. A Mixed user was defined as a user who performed at least one action that would have placed them in the Pro-Fake community and at least one action that would have placed them in the Anti-Fake community.

Captain Marvel From February 15 to March 15, 2019, we collected tweets for our analysis using Twitter's REST and Streaming APIs. We centered on the collection of all non-reply/non-retweet origin tweets that used #BoycottCaptainMarvel and #AlitaChallenge during our period of interest, or used "Alita" along with one of a set of keywords used in the contentious comic-book Twitter discussions (e.g. "SJW", "Feminazi"). We also collect all quotes, replies and retweets of these origins. Finally, we collected timelines of the central users and all non-reply/non-retweet origin tweets that provide information about the general Captain Marvel movie conversation. To use cross-platform information to understand community structure we collected author, subject, and viewership information for all YouTube video URLs shared through Twitter. In total, we collected approximately 11 million tweets.

We used a CASOS developed machine-learning tool [57] to classify the twitter users in our data as celebrities, news agencies, company accounts, and regular users. We used ORA [25], a dynamic network analysis tool, to create and visualize Topic Groups and Louvain clustering that was used to explore the Twitter communities that shared and responded to the misinformation

campaigns investigated. Topic Groups are constructed by using Louvain clustering on the intersection network between the Twitter user x Twitter user (all communication) network and the Twitter user x concept network. The resulting Topic Groups thus provide an estimation of which Twitter users in a data set are communicating with each other about the same issues.

2.4 Results

Our main analysis focuses on two different aspects of disinformation campaigns, distilled into two different case studies around the release of two controversial movies. We present the main results obtained for each of the studies, but for a complete description of each referer to the main publications [11, 12]. The *Black Panther* study involved comparing how each type of false tweet spread (retweeted) and was responded to (replies and quotes, and retweets of replies and quotes) through each community. We calculated the half-life (how long does it take for 50% of the retweet activity to happen) of the origin tweets themselves, the aggregate quotes related to each origin tweet and the aggregate replies to each origin tweet. We then compared these across the three communities described in the previous section. The *Captain Marvel* study presents an interesting opportunity to assess how the framing of a disinformation campaign, that otherwise originated in the same communities and with a similar presence of celebrity and news-like accounts, can affect the success in the diffusion achieved by each of them.

2.4.1 Speed of diffusion by origin post, response type, and communities in the Black Panther discussion

Community description Table 2.1 summarizes the total number of users in and total number of tweets by each of the three communities defined earlier. The Anti-Fake community is the largest and has the most activity, followed by the Mixed community. We further analyzed the behavior of the 2060 users in the Mixed community by examining the time trends of their activity. We checked for consistency in behavior by seeing whether the users went from supporting Pro-Fake posts to supporting Anti-Fake posts over time or vice versa or bounced back and forth in their support (indicating inconsistency). Table 2.2 summarizes the number of times Mixed users switched between Pro-Fake and Anti-Fake tweets (in either order).
Approximately 90% of users switched only once. Of those users that switched an odd number of times, 98% started by retweeting Pro-Fake Attack tweets and switched to retweeting Anti-Fake Attack tweets. A total of 271 user accounts from the Mixed community have been suspended and another 99 could no longer be found. Of the suspended accounts, 1 switched 7 times, 2 switched 4 times, 31 switched 3 times and 17 switched 2 times.

Diffusion of posts We compare the speed of diffusion of false Twitter stories across different story types, response types, and communities. One main observation is that the fastest of any tweets by at least an order of magnitude are some of the Anti-Fake quotes diffusing through the Anti-Fake and Mixed communities and the top Satire Attack story diffusing through the Anti-Fake community. For the top Fake Attack story (marked by a solid line in Fig. 2.1), the replies coming from the Anti-Fake community diffuse at speeds that are a magnitude above the speed

Community	# of Users	# of Tweets
Pro-Fake	1249	2147
Anti-Fake	47,016	116,541
Mixed	2060	35,916

Table 2.1: Community summaries

# of Switches	# of Users
1	1855
2	100
3	88
4	6
5	6
7	3
8	1
11	1

Table 2.2: Number of users who switched between supporting Pro and Anti-Fake Attack tweets

at which the origin post diffused in the Pro-Fake and Mixed communities. For at least the top 4 Fake Attack stories it appears that the quote responses diffuse at a speed as fast or greater than the origin posts themselves, implying that using quotes may be a helpful response option to false information if the intent is to rapidly meet the spread of the false information. An interesting difference between the Fake Attack and Fake Scene types of false stories is that while their top speed and total retweets are of similar magnitudes, most of the Fake Scene stories have slightly longer lifetimes than most of the Fake Attack stories and there is almost no response from the Anti-Fake community to the Fake Scene stories. This could be the case because there were fewer Fake Scene origin posts or because the Fake Scene stories were not reported in mass media outlets as the Fake Attack stories were. Whether the lack of response is related to the slightly longer lifetimes is unknown. Putting aside the top Fake Attack, which has some interesting timing associated with it, it appears that the Satire Attack stories also all have longer lifespans than the other types of stories.

Also, the speed of retweets of the Satire Attack posts and the responses to them more slowly decrease over the lifetime of the story than speed of retweets of Fake Attack or Fake Scene origin posts or Anti-Fake quotes attack those origin posts. This is the case even for Satire stories that do not reach as high a top speed of diffusion of Fake Attack or Fake Scene posts. If this is true of satire tweets in general, it may mean that while satirical responses don't go viral as quickly as the original fake story or as debunking quotes attacking those stories, they do appear to last longer. These attributes may mean satire responses are better suited as a response to longer-term false information campaigns rather than event driven viral ones. There is also almost no response to the Satire Attack stories from the Pro-Fake community. The mixed community (users that supported both Pro-Fake Attack and Anti-Fake Attack tweets) present an interesting group for discussion.

Previous work [9] reported that replies to Satire Attack posts included users who mistook the satire as Fake Attack posts, so confusion could have played a role. Since a large majority of the Mixed users switched from Pro-Fake to Anti-Fake actions it is possible that learning is the reason for many of the switches, but it is not possible to rule out the desire to create noise without further exploration of the timing of specific responses. It is important to note that even if most

Mixed users did learn that the Fake Attack and/or Fake Scene posts were indeed fake, they also are responsible for a larger amount of diffusion of those stories than their exclusively Pro-Fake counterparts.

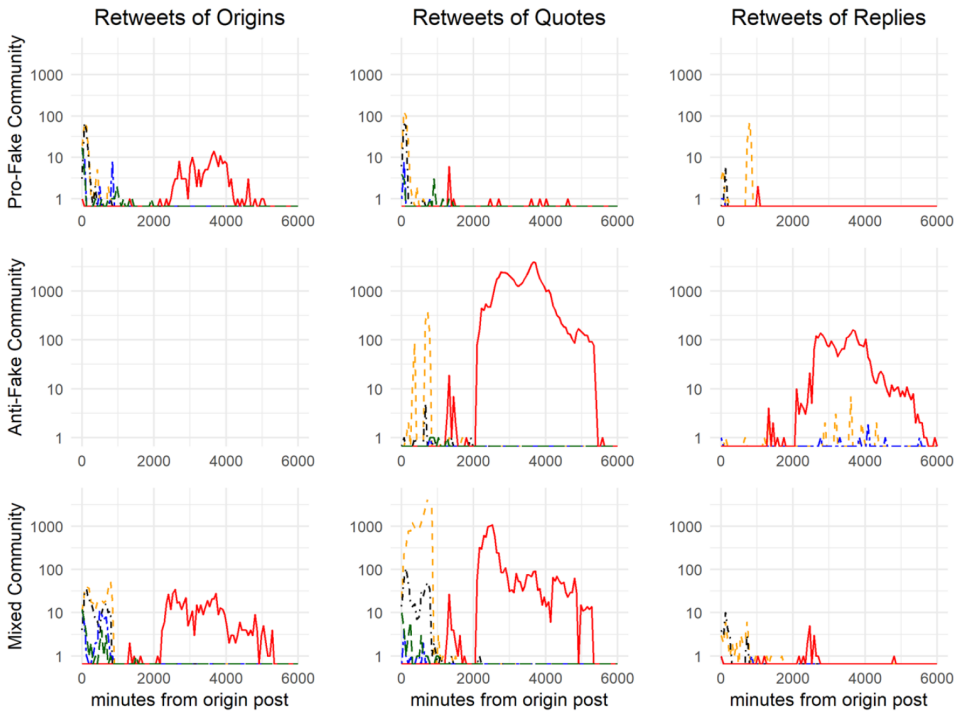

Figure 2.1: Retweets related to the top 5 Fake Attack origin tweets over time by community. Each color/linestyle represents the response to a different origin tweet. The speed of retweets is given per 5-min overlapping bins and is presented in a log-scale. For these stories, the quote and reply responses attacking the Fake Attack posts were spread much more successfully than the origin tweets. The Mixed community is a significant source of retweets of both origin tweets and responses to them. The top most spread Fake Attack story (solid line) shows somewhat different behavior than the others in that there is a lull in activity between the time of the origin tweet and the bulk of the retweets, quotes, and replies.

2.4.2 How framing affects the success of boycott campaigns in the Captain Marvel discussion

Diffusion on Twitter We compared the diffusion of the original boycott campaign tweets and responses to them over time. As shown in Figure 2.3, direct calls to boycott the Captain Marvel movie started more than a month before its official release on March 8, 2019, without gaining much traction on Twitter except for a day or two before the movie's release (green line). In contrast, during the same period there was an increase of discussion of "Alita" using harsh "culture war" phrases aimed at the *Captain Marvel* movie (Charged Alita). Most striking was the relatively rapid spread of #AlitaChallenge after March 4 (i.e. after the new use of it to attack

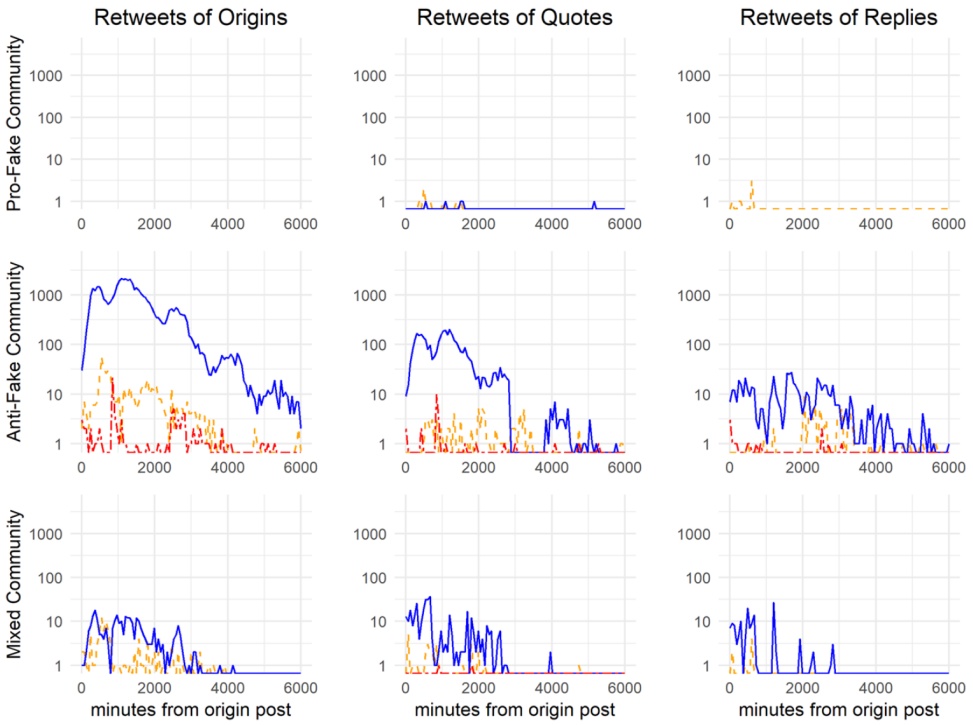

Figure 2.2: Retweets related to the top 3 Satire Attack origin tweets over time by community. Each color/line-style represents the response to a different origin tweet. The speed of retweets is given per 5-min overlapping bins and is presented in a log-scale. Unlike with the Fake Attack tweets shown in Fig. 2, here the retweets of the origin posts spread faster and to a larger extent than the quotes and replies that are responding to them. The quotes and replies are coming from within the same community (and are therefore of the origin tweets) while the response from the Pro-Fake community is almost non-existent.

Captain Marvel promoted by politically right-wing commentators). Overall, most support for these campaigns occurred between March 4 and March 12. Figure 2.3 also shows that most of the support for responses to the original campaign tweets occurred after the movie was released and, based on visual inspection, the majority of this support was for responses critical of the various boycott campaigns. The #AlitaChallenge campaign, while being the most directly supported, was also the most widely criticized, followed by responses to the Charged Alita conversations. It is noteworthy that the spike in negative responses coincided with the end of the majority of the support for #AlitaChallenge or Charged Alita tweets. This behavior is similar to what was previously noted in the Black Panther study, which is a similar event with contentious messages and debunked rumors. Tweets that directly pushed #BoycottCaptainMarvel or pushed both sets of hashtags were not responded to as great an extent.

Originating and Responding Twitter Communities In order to explore which Twitter communities in the overall Captain Marvel conversation the #AlitaChallenge and #BoycottCaptainMarvel campaigns originated and spread to, we calculated the Topic Groups and constructed the

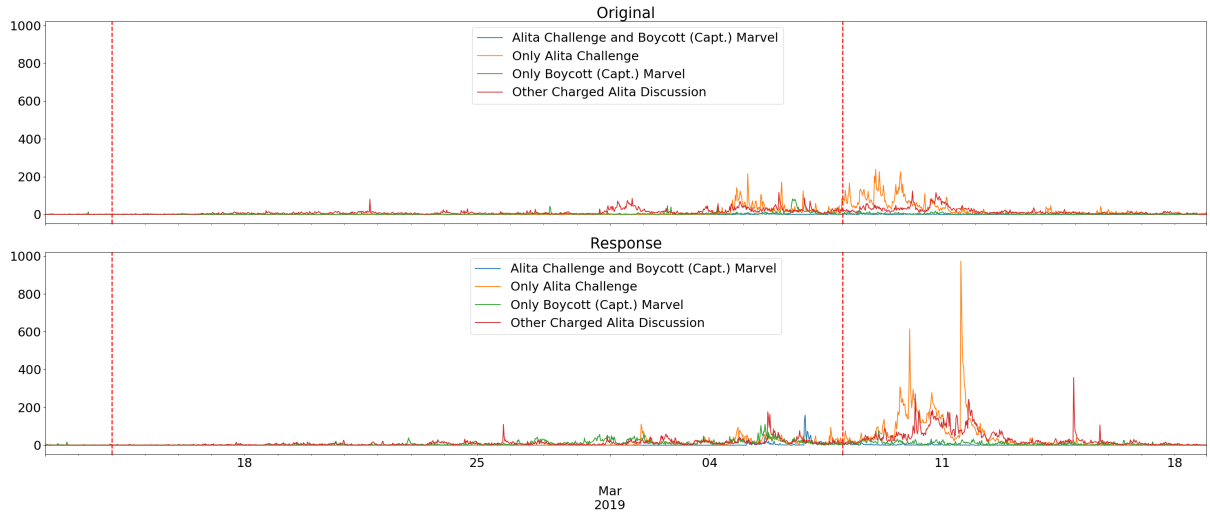

Figure 2.3: Diffusion of boycott/anti-Captain Marvel Alita campaigns. Retweets per 30-minute window of originals (top) and responses (bottom) for tweets that shared 1) both #AlitaChallenge and #BoycottCaptainMarvel, 2) only #AlitaChallenge, 3) only #BoycottCaptainMarvel, and 4) tweets with "Alita" and harsh words but no #AlitaChallenge. Vertical dashed lines are the opening of Alita: Battle Angel (leftmost) and Captain Marvel (rightmost)

Topic Group x Twitter user network. Prior to calculating the Topic Groups, we removed both main hashtags so that the resulting groups would not be based on them as inputs. Figure 2.4 shows that a significant majority of origin tweets came from users who were found to be in topic group 8 or 13 and that origin tweets for all four campaign types were found in the same topic groups. Topic Group 8 includes the account that originally high-jacked #AlitaChallenge as well as accounts and concepts associated with Comicsgate controversies. Topic Group 13 appears to a group more focused on comic book movies in general. Figure 2.4 also shows that the responses to the different boycott campaigns were more spread out among several more topic groups.

In addition to using twitter actions to communicate (quotes, replies, and retweets), URLs were also shared and helpful in exploring whether the four campaigns were being shared in the same communities. Using the YouTube data mentioned above, we constructed an Agent x Agent network where the nodes are Twitter user accounts and the links represent the number of shared YouTube authors. For example, if Twitter user A tweets out YouTube videos by author X and Y over the course of our collection period and Twitter user B tweets out videos by X but not Y then the link between A and B is 1 (because they both shared videos by author X). As links with weights of 1 may not be that indicative of broader community (as it could signal only one shared video in addition to one shared author), we examined the subset of this network were the minimum link weight was 2 shared authors as shown in Figure 2.4a and 2.4b.

Figure 2.5 shows that there is one main dense cluster of users and one smaller dense cluster. The main cluster was found by inspection to be Twitter accounts mostly involved with anti-Captain Marvel, anti-Marvel, right-wing politics, and "Red Pill" and Comicsgate-like misogeny or anti-diveristy sentiment. The smaller cluster (which is a separate louvain group as shown

in Figure 2.5b) appears to be Twitter users interested in comic book and other movies without an expressed political/cultural bent. Figure 2.5a shows that all four boycott related campaigns had origins and support from connected users in the main cluster (note that this cluster appears to have exisited prior to the release of Captain Marvel). It should also be pointed out that the Twitter user who began the hijacking of #AlitaChallenge is not central to the larger cluster though they are connected to the central cluster.

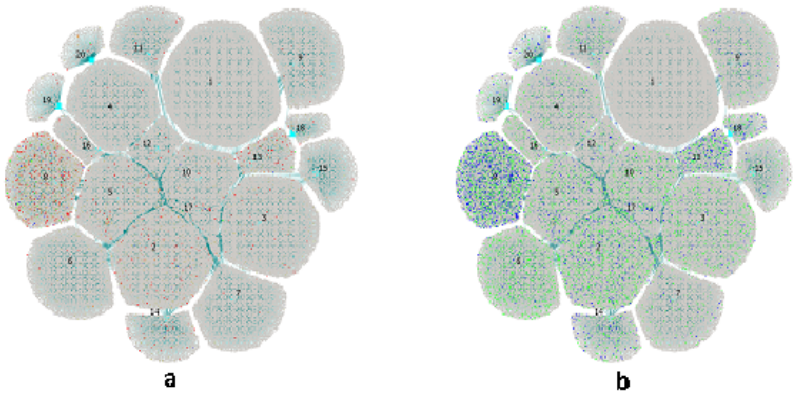

Figure 2.4: Twitter Topic Groups Network. Numbers and light blue squares represent Topic Groups. Colored round nodes in panel **a** represent origin tweets of the four different types. Colored nodes in panel **b** represent the origin tweets and retweets of origin tweets in blue and replies and retweets of replies in green.

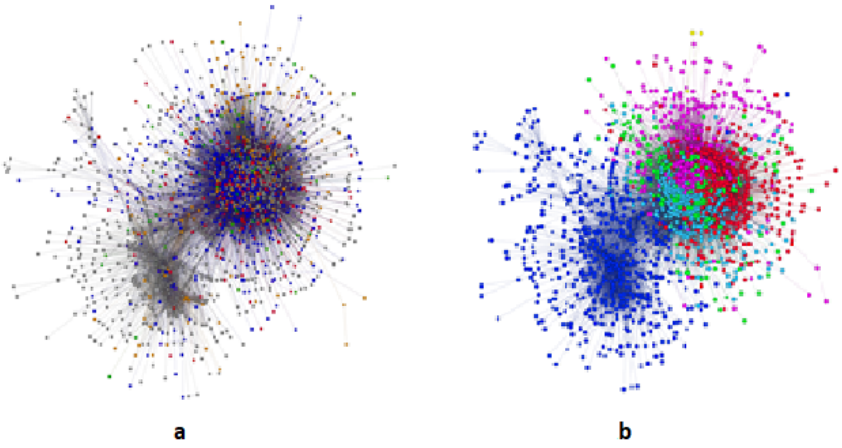

Figure 2.5: Twitter user x Twitter user shared Youtube author network. Link weights at least 2. In panel **a**, the four colors represent the four types of boycott-related campaigns. Panel **b** is colored by louvain grouping.

2.5 Limitations

The limitations of our analysis include the fact that in both cases we focused on the main tweets all within the same conversation context. The generalizability of the results should be tested using other false information datasets collected in different contexts. As mentioned, a closer look at the timing of false posts and responses across many Twitter contexts will help determine whether and how different responses affect the diffusion of false information (e.g. did the lack of response to the Fake Scene or Satire Attack stories allow them to live longer). Additionally, future work should include analyzing how different stories diffuse through communities that are not defined by their reaction to such stories. This will enable further analysis of which communities are more susceptible to starting or continuing the diffusion of false information in Twitter discussions.

2.6 Discussion

In this work we explored different characteristics of disinformation campaigns around the release of two Marvel Blockbusters that experienced significant amounts of contentious discussion around their release. In the context of the Black Panther release, we observed that debunking and mocking quote responses appear to diffuse faster in the community attacking false stories than the false stories themselves and supporting quotes do in the community promoting the false stories. Satirical responses appear to have longer lifetimes and, in some cases, higher speed of diffusion than other false stories. Our research also examined the importance and some attributes of those users who appear to act to both promote and attack the spread of false information. We hope that this work can help inform future research and decision making regarding the response to false information online. We also used the interesting case of the boycott-based-on-misinformation campaigns in the Captain Marvel twitter discussion to examine the origins, actors, and diffusion of misinformation. Overall, we found that the origins of all four types of campaigns were located in the same communities that share right-wing/anti-diversity sentiments through YouTube. We found that though the origins and discussion take place in similar communities and with a similar presence of celebrity and news-like accounts, the diffusion of the more indirect #AlitaChallenge and the more direct #BoycottCaptainMarvel occurred in different ways. It is noteworthy that in both case studies we found that the spike in negative responses coincided with the end of the majority of the support for each disinformation campaign.

Moreover, #AlitaChallenge and related discussions traveled to a much greater extent than #BoycottCaptainMarvel. This may have occurred because the #AlitaChallenge campaign on the surface presented a negative activity (boycotting Captain Marvel based on misinformation) as a positive one (supporting the strong female lead in Alita). Positive framing of actions has been found in past research to motivate participation, though not always. Our results may more strongly support the idea that #AlitaChallenge was more successful due to 1) successful hijacking of the hashtag by someone connected to the right-wing/anti-diversity comic book movie community and 2) the coordinated retweeting of a group of similar attempts to push #AlitaChallenge.

The purpose of this work should not be taken to be to help in the design of successful misinformation but rather to assist in the understanding of different methods intentionally used to promote false information. This may help with the design of effective and efficient community level interventions. Future work should delve more deeply into the cross-platform and cross-network nature of these conversations with an eye toward how that may improve our ability to classify the intent and effect of various campaigns.

Chapter 3

On Melding Network and Linguistic Polarization Methods: A Case Study on the 2019 South American Protests

3.1 Introduction

Political polarization is a widely-studied topic across multiple research disciplines [66, 98]. In computing and information systems literature involving social media data, two parallel lines of research – network-focused and language-focused – have made significant inroads into better understanding of complex political polarization in the era of ubiquitous internet. However, little or no literature exists melding these two research directions in a synergistic way. Our second observation on the current state of literature is the stark contrast between research attention on political polarization involving countries speaking English as first language (specifically the US) and countries speaking languages other than English. A major contributing factor to this attention difference is the sheer mismatch in natural language processing (NLP) resources between English and most other languages [91]. In this paper, we focus on a major political event *outside US* in a language *different from English* and show that the synergy between network-focused methods applied on news sharing patterns and language-focused methods can offer us a better understanding of the manifestation of polarization. We consider the 2019 South American protests, a major recent event that has received little or no attention from the computational social community. These protests started in Ecuador and were followed by Chile, Bolivia and Colombia and effectively paralyzed the countries for months. A central theme in all these protests was a massive online presence and the reported involvement of international and regional actors that sought to influence their evolution. While social media response during protests has received research attention, little or no publicly available data exists to further our understanding. Through this work, we release a data set of over 15 million tweets with weakly labeled stance of approximately 500k users. Via this substantial corpus, we analyze the online polarization during the South American protests along two dimensions: Polarization in language and in news sharing patterns. The contributions of this chapter are the following.

- ***Methods*:** We present a novel method to mine stance in Twitter conversations that requires minimal supervision and leverages users' endorsement of politicians' tweets and hashtag campaigns with defined stances towards the protest (for or against). The reliance on both signals provides different avenues to assess the robustness of the labels obtained and ameliorate the problem of hashtag hijacking. The mined stances are used to segregate the user pool into two groups: one in favor of the government, and the other, against it. The user stances and the subsequent grouping based on them, set up the basis for our study of the polarization observed in these networks.
  Our work relies on a domain expert who has comprehensive knowledge about the sociopolitical issues in all four countries involved. Having access to annotation expertise can be a double-edged sword. On one hand, the performance of machine learning systems benefits from superior annotation quality [4]. On the other hand, typically, expert annotators are costly. Our method relies on a weakly labelled approach and thus do not require a large number of labels. However, the risk of relying on automated methods for tasks with social impact is well-documented [56, 96]. *How do we validate our weakly labelled approach without requiring any further supervision?* To this end, we show that a recently proposed unsupervised method to quantify linguistic polarization [66] can be useful in a synergistic way. As we already mentioned, a well-known barrier to sophisticated linguistic analysis on language other than English is scarce computational linguistic resource availability. The machine-translation based linguistic polarization framework we use in this paper has only been previously used in US political contexts [66, 67]. We demonstrate that this method can be extended to a different language and can be used to characterize and quantify the linguistic polarization in countries not speaking English as the first language and experiencing a completely different sociopolitical crisis. We further validate these partitions by presenting strong evidence of polarization in news sharing and information diffusion by users, consistent with their stances towards the government. To the best of our knowledge, our work is one of the first few that combines network-based methods and language-based methods to jointly explore the nature of polarization in a user's news sharing behavior and her language usage.
- ***Social*:** We look at a globally important event: the series of South American protests that took place in multiple countries in 2019. We observe that linguistic polarization mainly manifested along ideological, political or protest-related lines. Moreover, we find strong evidence of the polarization in users' news sharing patterns, consistent with their stances towards the government. This can have pervasive effects on public discourse and political literacy.
- ***Resource*:** To the best of our knowledge, no large scale social media dataset relevant to protests spanning multiple countries exist. We hope that the rich network and semantic structure present in the data will be helpful not only for assessing polarization during these events but to advance stance classification efforts in a non-English language. The dataset is made publicly available at: `https://doi.org/10.5281/zenodo.6213032`.

3.2 Literature Review

Protests in the age of Social Media There is a long standing debate in political science concerning the role that media (social media in particular) plays in collective action during political unrest. Apart from the methods employed, the considered data sources also exhibit contrast. For instance, Dewey et al. [33] and Wolfsfeld et al. [131] draw contradicting conclusions from cross-sectional national surveys during the Arab Spring. The former found no significant correlation between social media use and mass protest [33]. On the other hand, Wolfsfeld et al. [131] reported that an increase in social media usage is much more likely to follow an increase in significant protest activity than to precede it. On the few studies on topics focused on the Latin American region, work has shown that usage of social media for political purposes significantly increases protest activities [121], and that individual attitudes of journalists were more predictive of their coverage than possible institutional pressures[93].

The second line of research focuses primarily on more stable democracies in wealthier countries and leverages medium to large scale social media data. This line of research can be further categorized into network-focused and language-focused studies. The former utilizes network-based methods centered around Twitter collections of protests to show that (1) protest communication networks are often fragmented and underutilized [51]; (2) peripheral participants play a critical role in the diffusion of protest messages [15]; and (3) the directionality of these networks can be leveraged to characterize the different roles played by individuals [16]. The latter focuses on textual features present on the social media posts [23, 44]. To our knowledge, little or no work has combined network-based and language-based methods to understand the nature of polarization. In what follows, we provide a brief overview of the research that has explored these separate, but related, dimensions of polarization.

Polarization in Language Political polarization has been widely studied across several disciplines with extensive focus on partisan US politics. Notable examples include studies on climate change [14, 42], gun control/rights [31], Supreme Court confirmations [29], economic decisions [90], congressional votes [97], polarization in media [100], etc. Recently, [66] presented a quantifiable framework for polarization that leverages machine translation (described in details in Section 5). Our use of this framework extends previous results in three key ways: we show that this method is generarizable to (1) different social media platforms (the machine translation-based method has only been applied to YouTube data primarily consisting of user-generated texts on news videos [65, 66]); (2) a language different from English; (3) and a completely different socio-political context of linguistic polarization during South-American protests.

Polarization in News consumption Much of the recent research on polarization in news consumption has centered around the cognitive dimension, while the social aspects have hardly been investigated [84]. Most echo chambers encountered empirically are a result of social media practices that exhibit polarized content engagement, rather than exposure [48]. However, there has been contradictory evidence in this regard, as studies have found limited evidence of challenge avoidance in online settings [34, 43, 49] while consistent evidence of reinforcement seeking. On a similar note, a review of numerous studies on Facebook showed that observed polarization is driven more by selective exposure resulting from confirmation bias than by filter

bubbles or echo chambers [112]. Polarization has also been identified as an important driver for diffusion of misinformation. This has been observed in the context of disinformation campaigns around the release of popular Marvel movies [11, 12] and in the diffusion of scientific, conspiracy theory and satiric Facebook articles [30]. This is not specific to Facebook, as the same relationship between polarization and confirmation bias has been observed on YouTube in a similar context [17]. In this work, we extend this observation to a different social media platform, namely Twitter.

Weak-Labeling of Social Media Data Utilizing weak signals in social media to train models has been explored in the past, achieving mixed results. In the field of ideology detection, several lines of work have focused on Twitter interactions to exploit a user's endorsement (given by retweeting, following, or liking behavior) of political figures to predict their general partisanship [50, 53, 132, 134], or to forecast the results of an election [115, 120]. In a similar vein, the usage of hashtag as weak label signals has been explored primarily in the field of stance classification, as observed on related SemEval [91] and IberEval [118] competitions. Works in this task found that using data from specific hashtags as additional training data could improve classification performance. However, this improvement was limited to certain topics partly because people use hashtags not only to explain their stance, but for many other purposes [40]. In this work, we build upon both approaches by leveraging hashtag usage of labeled political figures to reduce the risks of using non-informative hashtags. Moreover, as described in the following section, this offers our proposed methodology different levels of validation to assess the robustness of the labels obtained.

3.3 Weak Labeling Methodology

We propose a weak labeling methodology that requires minimal labeling effort and leverages users' endorsement of politicians' tweets and hashtag campaigns with defined stances towards the protest (for or against). The reliance on not only hashtags, but also endorsement of political figures provides different levels of validation to assess the robustness of the labels obtained. In particular, as we show in this section, we can leverage politician usage of the labeled hashtags to identify hashtags that were used to broadly by both groups. By excluding these instances, we can ameliorate the problem of hashtag hijacking that occurs with trending hashtags. We believe that this methodology holds promise for the development of large-scale databases for the analysis of similar contentious events (with the active involvement of local political figures).

Political Figures A set of 25 of the most prominent verified partisan actors for and against the protests were identified in each of the countries. These seed users included the most prominent members of each government involved with the protests (president, interior/defence minister, etc.). We also identified other leaders of the government's coalition party that were vocally supportive of the government (in Twitter). Similarly, we identified the movements/parties that were in opposition of the government (by promoting the protests) and included their most vocal/prominent members. We then compiled the list of their Twitter friends (people who they follow), ordered them by their number of followers, and labeled the stance of partisan actors.

These actors were classified as: politicians or political organizations, media figures or militants (influencers who describe themselves as partisan towards a politician or social movement).
The labeling procedure was as follows. A domain expert fluent in Spanish and with knowledge of the events that took place around each of the protests was presented the user card (including their description and location) of a labeling candidate and their timeline (tweets and retweets done by the user) in the period corresponding to the protest of the relevant country. Based on this information, the domain expert determines the type of user being considered and whether she was "in favor" (1), "against" (0), or if their stance was undetermined (-1) towards the government of the relevant country. In this process, 1,028 users are labeled by the domain expert. Table 3.1 summarizes the stance distribution of the political actors across different countries.

	Political Figures		**Stance-Tags**	
	Against	Pro	Against	Pro
Bolivia	83	31	180	91
Chile	164	183	452	254
Colombia	148	182	204	180
Ecuador	124	113	318	141

Table 3.1: Distribution of labeled political figures and stance hashtags.

Stance-Tags We construct a set of hashtags that occurred at the end of the body of a tweet and were consistently used in "in favor" or "against" the government of each country . We call these trailing hashtags "stance-tags". We label a first set of 778 hashtags as follows. Annotators fluent in Spanish and aware of the relevant events were presented tweets using these hashtags during the period corresponding to the protest of the relevant country. The annotators determine if the candidate hashtag was used primarily in tweets taking a position "in favor" (1), "against" (0), or neutral/undetermined (-1) towards the protests or government.
Next, we expand this set of hashtags by others that co-occurred exclusively with tweets for one side of the labeled set, yielding 838 extra hashtags. The validation of these second set was not as thorough as the first set and were only read to eliminate hashtags unrelated to the protest or deemed too general (in cases when this was not apparent a small sample of tweets were analyzed as described before). We finally augment these with any missing hashtags used exclusively by labeled political figures from one side of the argument. As before, this set was only validated by a cursory analysis to remove any hashtag unrelated to the protests. This augmentation step added 227 new hashtags, and the final distribution is presented in Table 3.1.

3.3.1 Stance-Tags Validation

Our weak labeling methodology relies on the hypothesis that users are more likely to tweet (or retweet) hashtags or political figures that are aligned with their stances during these events. Hence, weak-stance labels are assigned to a user if their percentage of tweets with a consistent stance-tag is above a given threshold. To test this hypothesis, we apply our methodology (just

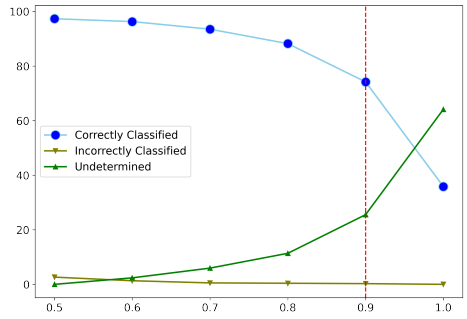

Figure 3.1: Performance of the weak labeling methodology on the labeled political figures at different probability thresholds. The chosen threshold for the construction of the dataset is indicated by the dashed vertical line.

based on stance-tags) to predict the stance of the political figures labeled. We can also use this exercise to determine a suitable threshold for the stance assignment. We limit our analysis to the 88.1% of labeled users that tweeted (or retweeted) at least 5 tweets containing a stance stag. We also present results excluding the set of extra 229 stance-tags obtained using this set of users in order to have a better assessment of the performance in the wild. Figure 3.1 presents the accuracy of the methodology at different probability thresholds. As expected, higher thresholds are more conservative in the assignment of a label (the percentage of undetermined users increases) but also decrease the likelihood of missclassification. However, in the most aggressive classification threshold, only 2.6% of the users are missclassified, which supports our starting hypothesis.
For the construction of the dataset released in this work, we opt for a conservative 90% threshold, which results in 74.2% correctly classified users but only a 0.3% (2 users) classification error. The reason for this conservative approach is that our validation set is comprised of highly political users, which could result in a higher likelihood of missclassification among more casual users [40].
Nonetheless, we are able to considerably increase the performance of this methodology (with the 90% threshold) by first including the aforementioned 229 hashtags, used exclusively by user of each side, which improves the accuracy to 80.0%. Lastly, we prune our hashtag set by removing tags that were used too frequently by users of a different stance. This results in the removal of 46 hashtags and brings the final classification accuracy of our proposed weak-labeling methodology to 88.6% of the hand-labeled political users.

3.3.2 Determining User Stances

We assign the User stance based on how prominently they tweet (or retweet) a stance tag or retweet a labeled political figure. The threshold used to determine the stance was obtained during the hashtag validation procedure described above and set at 90%.

Hashtag usage Users were assigned a stance if they used stance tags either in their tweets (or retweets) or in their user description. In both cases, a stance was assigned to a tweet (or description) if it contains hashtags with the same stance, otherwise it was deemed inconsistent. As before, we only proceeded with users that had at least 5 tweets with a consistent stance or if at least one description was consistent. As less than 1% of labeled users were labeled based on their descriptions, we do not desegregate results based on the origin of the label. A stance was assigned to a user if at least 90% of their tweets had the same stance. The number of users classified and their distribution is presented in Table 3.2.

	Pro	**Against**	**Inconsistent**	**Total**
Bolivia	46 040	57 654	2 717	106 411
Chile	20 478	220 648	381	241 507
Colombia	16 460	55 883	154	72 497
Ecuador	18 909	48 478	230	67 617

Table 3.2: Government Stance of users based on hashtag usage.

Endorsement of Political Figures The procedure followed to assign a stance to a user based on their endorsement of political figures, follows the same logic as before. As such, users were assigned a stance if at least 90% of their retweets of labeled political figures are from users with the same stance. As before, we only proceed with users that had at least 5 retweets of these users. The number of users classified and their distribution is presented in Table 3.3.

	Pro	**Against**	**Total**
Bolivia	29 573	17 444	47 017
Chile	29 548	78 735	108 283
Colombia	22 165	60 082	82 247
Ecuador	20 957	32 243	53 200

Table 3.3: Government Stance of users based on endorsement of political figures.

The final user stance was determined by combining the labels obtained by both methodologies, allowing us to validate the stances assigned to the users. There was a 41.9% overlap of users labeled by both methodologies, while 46.8% of the users were labeled based on their hashtags and 11.3% only based on their endorsement of political figures. Whenever there was an inconsistency between the label assigned by both methodologies (or if one method assigned an inconsistent label for a user), we deemed that label inconsistent. Table 3.4, presents the stance distribution of the labels among users matched by both methodologies. Importantly, in the worst case (Bolivia) a 10.6% of the users had an inconsistent stance, which provides evidence for the robustness of the labels constructed.

The final distribution of the consistent weakly-labeled users and the number of original tweets (not including retweets) posted by them is presented in Table 3.5. We also include counts for

	Pro (%)	Against (%)	Inconsistent (%)
Bolivia	52.31	37.06	10.62
Chile	17.33	76.39	6.27
Colombia	23.23	71.49	5.28
Ecuador	32.91	62.12	4.97

Table 3.4: Stance Distribution of users matched by both methodologies

users in the two-hop neighborhood of labeled users (these are also part of the released dataset), as we expect that the rich network structure present in the data will be helpful not only for assessing polarization during the event but to advance stance classification efforts in languages other than English.

		Bolivia	Chile	Colombia	Ecuador
Against	Users	58 727	221 641	79 908	51 545
	Tweets	2 079 286	7 648 773	2 258 276	1 248 620
Pro	Users	54 776	33 327	28 310	25 566
	Tweets	1 447 120	2 167 763	1 164 129	493 509
Neighbors	Users	668 815	860 824	556 821	457 241
	Tweets	7 803 498	8 518 147	4 978 669	4 307 034

Table 3.5: Number of Weakly-Labeled Users and their original tweets (not including retweets). We also include the corresponding counts for users in the two-hop neighborhood of labeled users.

3.3.3 Ethical Considerations

We make our data publicly available[1] and, to adhere to Twitter's terms and conditions for sharing data, we do not share the full JSON of the collected tweets. Instead, we provide their respective tweet or user IDs, the type of tweet (Original, Reply or Quote), and in the case of weakly labeled users or tweets, their assigned label. Since the Tweets will have to be re-hydrated, if a user deletes a tweet (or their account), it will not be available for analysis ensuring that the user's *right to be forgotten* is preserved. However, for the hand-labeled political figures, given their public role during these events, we not only provide their user ID but also their user name and user type (as described above). We also release the full set of labeled stance tags.

3.4 Linguistic Polarization

We employ a recently-proposed framework [66] to quantify linguistic polarization in large-scale text discussion data sets. In the original paper, the authors applied this framework to English

[1] `https://doi.org/10.5281/zenodo.6213032`

YouTube comments in the context of polarization in US cable news networks viewerships. Our results indicate that the methodology is generalizable to discussions on a different political crisis, in a different language (Spanish), and manifested in a different social media platform (Twitter). This framework assumes that two sub-communities (e.g., the sub-community favoring the protest and the sub-community opposing the protest) are speaking in two different *languages* (say, $\mathcal{L}_{pro}$ and $\mathcal{L}_{against}$) and obtains single-word translations using a well-known machine translation algorithm [109]. Since $\mathcal{L}_{pro}$ and $\mathcal{L}_{against}$ are both in fact Spanish, ideally, any word w in $\mathcal{L}_{pro}$ should translate to itself in $\mathcal{L}_{against}$. However, when a word w_1 in one language translates to a different word w_2 in another, it indicates w_1 and w_2 are used in dissimilar contexts across these two *languages* signalling (possible) disagreement. These disagreed pairs present a quantifiable measure to compute differences between large scale corpora as greater the number of disagreed pairs the farther two sub-communities are. A formal description follows.

Let our goal be to compute the similarity between two languages, $\mathcal{L}_{source}$ and $\mathcal{L}_{target}$, with vocabularies $\mathcal{V}_{source}$ and $\mathcal{V}_{target}$, respectively. Let $translate(w)^{\mathcal{L}_{source} \rightarrow \mathcal{L}_{target}}$ denote a single word translation of $w \in \mathcal{V}_{source}$ from $\mathcal{L}_{source}$ to $\mathcal{L}_{target}$. The similarity measure between two languages along a given translation direction computes the fraction of words in $\mathcal{V}_{source}$ that translates to itself, i.e.,
$Similarity(\mathcal{L}_{source}, \mathcal{L}_{target}) = \frac{\Sigma_{w \in \mathcal{V}_{source}} I(translate(w)^{\mathcal{L}_{source} \rightarrow \mathcal{L}_{target}} = w)}{|\mathcal{V}_{source}|}$. The indicator function returns 1 if the word translates to itself and 0 otherwise. The larger the value of *Similarity* $(\mathcal{L}_{source}, \mathcal{L}_{target})$, the greater is the similarity between a language pair.

We constructed $\mathcal{L}_{pro}$ and $\mathcal{L}_{against}$ by combining all the main tweets by a user of a given stance (this includes tweets not related to the protests) ensuring that a retweeted tweet is included only once. Following [66], for each of constructed corpora, we trained a 100-dimensional FastText embedding [19], with basic pre-processing that included removing URLs, mentions and punctuation. Next, using a well-known machine translation algorithm, we translate the top 10k words in one language to the other and examine disparities. We followed the same experimental protocols as [66] and (1) used stop-words as anchors for the translation as these are the most likely to maintain their meaning across the different groups; and (2) given the imbalanced size of the two corpora, we sub-sampled the majority community (against the government) to match the size of the smallest community across the countries (Ecuador's Pro-Government community). This guaranties a fair comparison between the two communities.

How can we linguistically validate our weakly labelled user stances? Before we delve deep into the qualitative and quantitative results of this machine translation-based framework, we first highlight how linguistic polarization can provide corroborating evidence and validate weakly labelled user stances. We contrast this linguistic polarization measure against a baseline similarity metric calculated by repeating the process with randomized splits (of the same size) for each of the countries. Our intuition is a randomized split will exhibit lesser linguistic polarization than a split guided by stance. Table 3.6 indeed shows that is the case. This results lends further credence to our method to determine stance and indicates how these methods can be applied in a synergistic way. We further note that the largest polarization was observed in Bolivia and Ecuador, while the lowest was observed in Chile (this difference being significant at a 95% threshold).

We next contextualize the observed linguistic polarization by highlighting notable examples of the mistranslated word pairs. Due to space constraints, in Table 3.7 we present examples only

	Stance-Split (%)	Randomized-Split (%)
Bolivia	57.11 (1.31)	93.37 (0.61)
Chile	68.21 (1.46)	91.37 (0.62)
Colombia	64.22 (1.13)	91.94 (0.67)
Ecuador	56.50 (1.12)	93.52 (0.53)

Table 3.6: Pairwise similarity between languages computed for: Pro and Against government communities (Stance-Split) and a baseline obtained from two random samples from the combined corpus (Randomized-Split). All similarity metrics are constructed with corpora of approximately the same size. We repeated the sub-sampling process six different times and report the mean and standard deviation (in parenthesis). The evaluation set is computed by concatenating the corpora and taking the top 5K words ranked by frequency.

for Bolivia and Ecuador (the results are consistent across all the countries studied). We find a polarization in language, mainly manifested along ideological, political and protest-related lines. Terms related to left-leaning ideologies in one community tend to be discussed in similar contexts as right-leaning terms (e.g. Socialism mistranslates to Fascism); terms related to law and order in one group are discussed in a similar context as the other discusses oppression (an informal term for police mistranslates to bastard or vandals to protesters). Importantly, we find that the motivations for the protests mistranslate to each other. For example, in the case of Ecuador, Decreto883[2] mistranslates to derogate which was one of the calls of the protests movements (a similar pattern is shown for Bolivia with the "overthrow" term). Finally, opposition leaders are discussed in similar context as government representatives.

[2]This refers to the decree 883 which proposed austerity measures and started the protests in the country

Country	**Theme**	L_{pro} (translation)	$L_{against}$ (translation)
Ecuador	Idiological	Fascistas (fascists)	Comunistas (communists)
		Socialismo (socialism)	Fascismo (fascism)
		Neoliberal (neoliberal)	Progresista (progressive)
	Protest	Chapas (police -informal-)	Infelices (Bastard)
		Vandals (vandals)	Protestantes (protesters)
		Decreto883 (name of policy)	Derogatoria (derogate)
	Political	Jarrin (Minister of the Interior)	CONAIE (Protest leaders)
		Matraca (supportive politician)	Innombrable (opposition politician)
		Lasso (supportive politician)	Correa (opposition politician)
Bolivia	Idiological	Comunismo (communism)	Neoliberalismo (neoliberalism)
		Socialismo (socialism)	liberalismo (liberalism)
		Revolucionario (revolutionary)	Republicano (republican)
	Protest	Autogolpe (self-coup)	Golpe (coup)
		Derrocar (overthrow)	Derrotar (defeat)
		Masistas (supporter of president's party)	Maleantes (malefactors)
	Political	evomorales (ex-president)	Criminal (criminal)
		Almagro (OEA's general secretary)	Delincuente (Delinquent)
		Evo (president's name)	Asesino (assassin)

Table 3.7: Notable instances of linguistic polarization by topic for Ecuador and Bolivia.

$\mathcal{L}_{\text{pro}}$	Bolivia	Chile	Colombia	Ecuador
Bolivia	-	51.98 (1.47)	46.00 (0.70)	42.98 (0.91)
Chile	51.98 (1.47)	-	69.22 (0.82)	54.42 (0.74)
Colombia	46.00 (0.70)	69.22 (0.82)	-	51.64 (0.57)
Ecuador	42.98 (0.91)	54.42 (0.74)	51.64 (0.57)	-

$\mathcal{L}_{\text{against}}$	Bolivia	Chile	Colombia	Ecuador
Bolivia	-	51.70 (0.47)	50.10 (1.2)	50.74 (1.13)
Chile	51.70 (0.47)	-	62.96 (0.64)	65.48 (1.72)
Colombia	50.10 (1.2)	62.96 (0.64)	-	57.58 (0.93)
Ecuador	50.74 (1.13)	65.48 (1.72)	57.58 (0.93)	-

Table 3.8: Pairwise similarity between languages computed for: Top) pro-protest communities in different countries, and Bottom) against-protest communities in different countries. All similarity metrics are constructed with corpus of approximately the same size. Standard deviations are presented in parenthesis. The evaluation set, is computed by concatenating the corpora and taking the top 5K words ranked by frequency.

Moreover, to test the robustness of the methodology to differences in dialects, we apply it to compare stances between the countries. The middle and bottom sections of Table 3.8 include the pairwise similarity of languages for pro- and against-government communities between countries. Note that, for both stance types, the language similarity is significantly lower for pairs that contrast Bolivia with any other country. This is consistent with the fact that Bolivia is the only country which has a left-leaning government and hence right-leaning protesters. Moreover, even with the added noise of local colloquialisms, we are still able to recover the ideological polarization between protest movements, when compared to Bolivia ("neoliberalism" still mistranslates to "communism" or "socialist" to "nazi"). These differences do not exist when comparing protests movements with similar ideological motivations (e.g. "socialist" in Ecuador translates to "socialist" in Chile). Importantly, we can use this methodology to mine knowledge from the corpora, as we find that local political leaders from the protest in one country mistranslate to their counterparts in the other country (e.g. "Correa" in Ecuador mistranslates to "Petro" in Colombia – an opposition leader).

3.5 Polarization in News Sharing Behavior

We started with a dataset of news agencies and journalist for the countries explored (this was obtained from the NetMapper software [3]). It had several limitations and was expanded by searching for the most important news agencies operating in each country, manually checking who they follow and adding agencies that were not included. This resulted in a list of 853 news agencies (or

[3] https://netanomics.com/netmapper-government-commercial-version/

major reporters) detailing their Twitter handles and main URL (if available). Notably, the list included agencies from Venezuela and Russia that predominately operate in the region, this is important as we explore influence campaigns on the protests. We then proceeded to identify the agencies that were either directly retweeted by a user or that had a user tweet/retweet a URL corresponding to their domain. The number of news agencies from each country resulting from this process is shown in Table 3.9.

	Bolivia	Chile	Colombia	Ecuador	*Regional
# Agencies	52	95	69	99	69

Table 3.9: Number of news agencies in each country. *The regional category includes regional Venezuelan and Russian media among others.

Filtering Irrelevant Media Tweets News articles identified in our data set cover topics ranging from the protests to sports. When studying the polarization of news consumption during the political event, it is important to first remove tweets which are irrelevant to the protests. It is not obvious if a tweet from a news agency is relevant or not, but many tweets in our data set contain the URL of an article that they reference. For this reason, we determined the relevance to the protest of a small set of the 900 most tweeted URLs in our dataset distruted among the different countries. We complemented this dataset with an additional set of URLs labeled by extracting subsection metadata from them. If the subsection referenced sports, culture or technology, the URL was labeled as irrelevant to the protests. Then, we assigned the URL label to any tweet that used it. The final sample distribution are presented in Table 3.10. We note that even though we are able to assign a label to more than 100k tweets, most of them contained duplicated text (as news media tend to tweet the same thing multiple times). The classification was done with the unduplicated dataset.
To classify the relevance of the tweets, we built a CNN text classifier [68] using 300 dimensional FastText embeddings trained on the combined datasets (both by stance and country) used to analyze the language polarization. We used 100 filters on 3 layers with filter sizes 3, 4 and 5 and a dropout rate of 50%. We achieved an accuracy and F1-score of 92% in a held-out test set. After predicting the labels of tweets (relevant or irrelavant to the protests), we obtain a dataset of 1,024,166 relevant and 675,496 irrelevant tweets. The distribution of the data set is shown in Table 3.10. The analysis of polarization in news consumption patterns presented in this works was done only on the tweets that are relevant to the protests.

Community Detection Among News Agencies After filtering out the irrelevant tweets, we keep the 383 subset of news outlets that are either directly retweeted by a user or that have a user tweet/retweet a URL corresponding to their domain (see Table 3.9). We then explore the community structure of these news outlets based on the homogeneity of their user bases. We define the bipartite network $\mathcal{G} = (\mathcal{N}, \mathcal{U}, \mathcal{E})$ where $\mathcal{N}$ is the set of news outlets; $\mathcal{U}$ is the set of users who have retweeted the posts of news outlets in $\mathcal{N}$, or tweeted (retweeted) a URL corresponding to one of the news agencies.

	Labeled Tweets (#)		Predicted (#)	
	Total	Deduplicated	Tweets	Users
Relevant	78 318	7 177	1 024 166	276 754
Not Relevant	53 671	8 156	675 496	247 324

Table 3.10: Distribution of the labeled tweets and resulting predictions after classification.

The bipartite graph, $\mathcal{G}$, can be described as a matrix $\mathcal{M}$ for which

$$\mathcal{M}_{i,j} = \begin{cases} 1, & \text{user } j \text{ shared (retweeted) news agency } i \\ 0, & \text{otherwise.} \end{cases}$$

We then define the co-occurrence matrix $\mathcal{C}^N = \mathcal{M}\mathcal{M}^T$, which counts the number of users shared by two news agencies in $\mathcal{N}$. As shown in Fig 3.2, media outlets are divided into three clusters: a) major local media (blue); b) regional Russian and Venezuelan media (red), including local left-leaning media; and c) local Venezuela media (gray). Note that, outlets are not only clustered geographically, but also ideologically as the local media cluster includes the major media organizations in each country, which, as mentioned before, were reportedly more likely to support the local governments (with the exception of Bolivia where the situation is reversed). Our clustering results support the hypothesis that media with similar stances are more likely to share homogeneous user bases providing evidence for confirmation bias as a possible driver for users' news diffusion behavior. In what follows, we present a detailed treatment focusing on Colombia, Ecuador, Bolivia, and Chile.

3.5.1 Quantifying the Polarization

We next explore the diversity in the user bases of the news outlets in a given region to estimate the polarization in news sharing by users with different stances. To this end, we compute the ratio of retweets from pro-government users for the regional Russian and Venezuelan news media clusters. We find that in Chile, Colombia and Ecuador, 90% of these outlets have 10% or less of the diffusion of their articles coming from pro-government users. Of the remaining 10% of these outlets, for the vast majority, the relevance of this user base was less than 20%. Whereas in Bolivia, the reverse is observed, as almost all news outlets in this cluster had at least 80% of their retweets coming form pro-government users (there was only one outlet that had 70% of this user base). This is consistent with the political orientation of the media organizations in this cluster and their documented support of the Bolivian government [54, 104], while Ecuador, Chile and Colombia have more right-leaning governments. The results provide further evidence of polarization in news sharing behavior, as users are more likely to retweet news aligning with their stances.

We further explore if the observed level of polarization can be accounted by the asymmetry in the user stance distribution found in each country. For each news outlet in the two clusters analyzed,

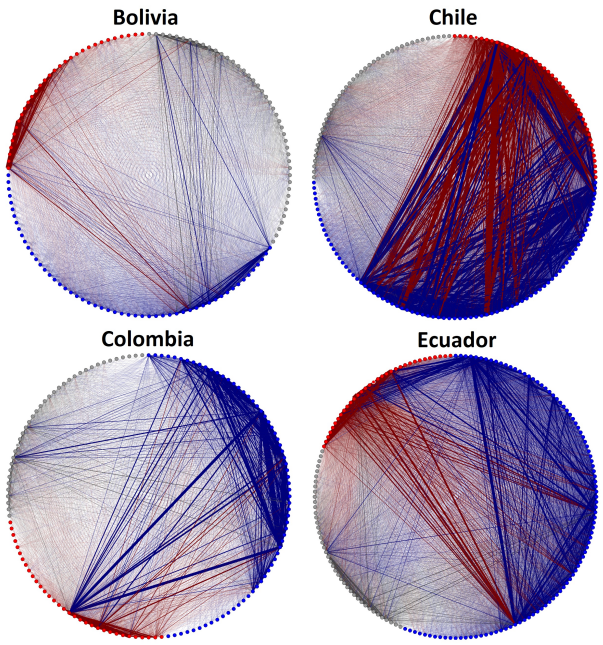

Figure 3.2: The clusters of the news outlets. The colors indicate different communities, where a) blue represents local media; b) red represents regional Russian and Venezuelan Media; and c) green represents local Venezuelan media. Edges connecting two dots (news outlets) represent common users that are shared by the two news outlets.

we compute the relative entropy of their user bases. For each news agency n, its relative entropy is defined as follows:

$$H(n) = -p_{pro} * log\frac{p_{pro}}{g_{pro}} - (1 - p_{pro}) * log\frac{1 - p_{pro}}{1 - g_{pro}} \tag{3.1}$$

where p_{pro} is the ratio of retweets for news media n from pro-government users and g_{pro} is the overall ratio of pro-government users in that country. The relative entropy $H(n)$ evaluates how p_{pro} differs from g_{pro} - the lower the value of $H(n)$, the more disproportional the level of polarization is with respect to the asymmetries observed in the stance distribution. The maximum possible value 0 of $H(n)$ is obtained when a news organization's user base matches the the stance distribution for the country. Fig 3.3 shows, for each of the countries studied, the distribution of relative entropy for news media in the clusters of local media and regional Russian, Venezuelan media. We observe that for all countries, regional media from Russia and Venezuela are disproportionately polarized. Moreover, in the case of Bolivia, we observe the highest level of relative polarization, an observation that is consistent with the language polarization levels described in the previous section.

3.5.2 Polarization through News Media Transitions

We finally analyze user transitions among the different clusters of news media. We adapt a methodology previously used to examine the transition behavior of users browsing websites re-

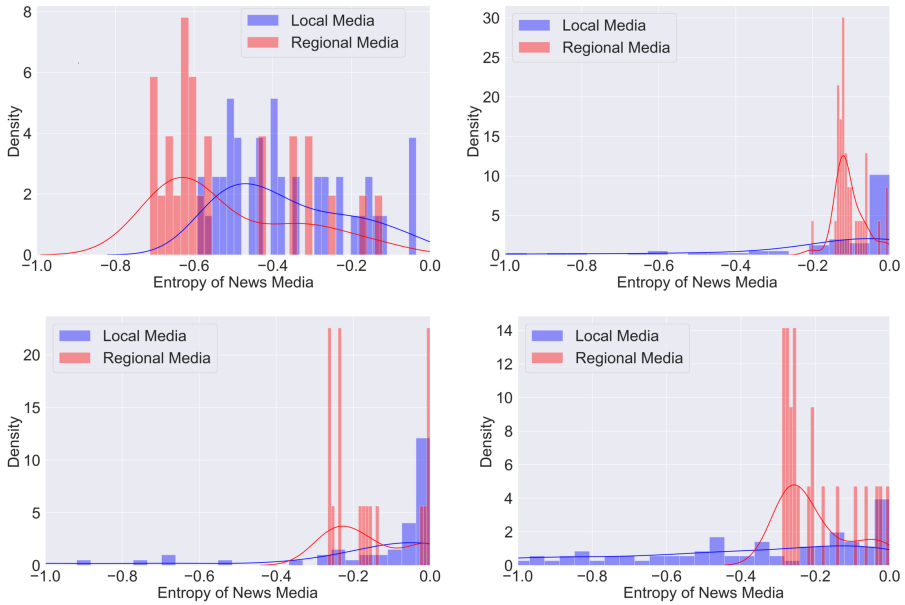

Figure 3.3: Distribution of relative entropy for local (blue cluster) and regional Russian and Venezuelan media (red cluster) in Bolivia (top left), Chile (top right), Colombia (bottom left) and Ecuador (bottom right).

lated to the topic of gun rights/control [73]. Specifically, we seek insights about how the political orientation of news media influences the type of news outlets that a user will retweet next[4] after having retweeted news supporting or objecting the protest. For each user, we represent her retweeting history as a Markov chain of the cluster of outlets the retweeted news come from. Then, we describe the distribution of the transition probabilities by an n-state transition matrix P_n, with elements $p_{ij} = \text{Prob}(X_{t+1} = j | X_t = i)$. We note that the row-wise sums are equal to 1. State 0 represents when the user shares news articles from the local news cluster, and state 1 from the cluster of regional Russian and Venezuelan media. Table 3.11 summarizes the transition matrix for Bolivia, Chile, Colombia and Ecuador.

To analyze the underlying trends of these matrices, we employ Summary Mobility Indices, which have been widely used in economics and sociology. They describe the direction of the mobility in the following way:

- Immobility Ratio: IR =$\sum_{i=1}^{n} p_{ii}/n$
- Moving Up (Left): MU (ML) = $\sum_{i<j} p_{ij}/n$
- Moving Down (Right): MD (MR) = $\sum_{i>j} p_{ij}/n$

where n is the number of states. As mentioned previously, the cluster of Russian and Venezuelan media is comprised of left-leaning outlets, including smaller media operations local to each country. We therefore rename the moving up index (from local media to Russian, Venezuelan media) as moving left (ideologically), and the moving down index as moving right (ideologically).

[4]An outlet can be retweeted either directly or indirectly via a tweet originating from their account, or a third party tweet containing a url with their domain.

Country		Local media	Regional media
Bolivia	Local media	95.26%	4.74%
	Regional media	3.67%	96.33%
Chile	Local media	64.35%	35.65%
	Regional media	21.13%	78.87%
Colombia	Local media	91.4%	8.6%
	Regional media	18.01%	81.99%
Ecuador	Local media	80.43%	19.57%
	Regional media	7.59 %	92.41%

Table 3.11: Transition matrix for Bolivia, Chile, Colombia and Ecuador

Country	IR	ML	MR
Bolivia	95.87%	2.05%	2.08%
Chile	73.5%	13.19%	13.31%
Colombia	88.38%	5.84%	5.78%
Ecuador	89.21%	5.24%	5.55%

Table 3.12: Summary Mobility Indices

The Summary Mobility Indices for different countries are included in Table 3.12. We note that the majority transitions are to the same state, which indicates users tend to share articles from the same community where the stances of the information align with their own. The likelihood of transitioning out of a user's community is generally low. In this regard, the transition matrices suggests (1) presence of eco-chambers in the midst of the protests and (2) confirmation bias in the way users choose to share content. Moreover, the observed polarization levels in news transitions are consistent with what was described for the language polarization in the previous section, with Chile showing the lowest level of polarization and Bolivia the highest.

3.6 Limitations

Our proposed weak-labeling methodology is not able to determine users with neutral stances towards the government (or the protests). This limits the scope of the analysis presented in this work to users with two defined stances towards the event: for or against. Characterizing the neutral users and estimating their prevalence in these events merit deeper exploration as they can potentially serve as the bridge between polarized communities. Also, the framework we used to assess linguistic polarization relies on word-level mistranslations. Contrasting more complex political beliefs expressed at the sentence level could be a worthy future research challenge. Finally, even though our findings suggest that polarization in news sharing is consistent with users' stances, we are not able to ascertain if this is due to confirmation bias in part of the users,

or a result of their exposure to said content because of filter bubbles.

3.7 Discussion

In this work we explore polarization in user behavior based on their stance towards the government during the 2019 South American protests. We employ a novel method that requires minimal supervision to label the stance of users. We make our resulting dataset publicly available. We focus on polarization in language and in news media sharing and show that together, the analyses shed vital insights. Our linguistic polarization results indicate that polarization largely manifests along ideological, political and protest-related lines; and show that we we can obtain interesting insights by inspecting the mistranslated pairs. We also find strong evidence of polarization in news sharing and information diffusion by users, consistent with their stances towards the government. The news media in our data set clusters with the political stances of their content. We find consistent evidence of polarization in the way users choose to share news on Twitter, as users tend to stay in the community of news media that shares information they are more likely to agree with. Moreover, we show the important role that regional Russian and Venezuelan news outlets like RT en Español and TeleSUR, played in the social media discussion of the protests throughout the region. This shows how effective these outlets have been in gathering an audience of left-leaning users in the region, an initiative that has been identified by other studies of these news outlets [54, 104]. Finally, we observe that along both dimensions of polarization explored, we obtain consistent results, with Chile showing the lowest level of polarization and Bolivia the highest.

Chapter 4

Protest Stance Detection: Leveraging heterogeneous user interactions for extrapolation in out-of-sample country contexts

4.1 Introduction

Public opinion towards governments and their policies have widespread social implications. Public support is often viewed as a kind of currency in political science: more public support can allow governments more freedom in their actions, whereas widespread public disapproval can constrain the set of actions that a government can take. Studies have found public opinion to be a strong predictor of policy [24, 78], and information about public opinion can impact the decisions and planning of politicians, businesses, foreign governments, and people. It is therefore unsurprising that global election and opinion polling has grown into a 6.78 billion dollar industry [1]. However, opinion polling often lags behind real events, and in unstable political situations, this can engender widespread uncertainty. Surveys, especially in dynamic or unstable political situations, are constricted by resources, and the diversity of participants is confined by the survey method. However, in some cases, large-scale Twitter data can serve as an alternative way to gauge public opinion [120].

Automated mining of social media data for the extraction of public opinion is a stance detection problem. However, users engage with one another through these platforms in a variety of ways: they can post original messages, reply to other users, or share the each other's posts. It has long been recognized that considering only an isolated sentence will provide an incomplete assessment of a user's stance towards a predefined target [37]. Despite this, little research has been done to integrate the different context levels available into stance detection models and, more importantly, to evaluate the impact these additions can have on classification performance [75].

In this work we explore the in-sample performance and generalizability of country-level models trained at various levels of context. We explore leveraging context at the level of single tweets, users, and social networks. For this task, we use a weakly-labeled large-scale Twitter dataset that

encompasses the widespread 2019 protests that erupted in Ecuador, Chile, Bolivia, and Colombia [125]. The regional nature of the dataset presents the opportunity to test the generalization capabilities of our proposed models in out-of-sample country data at different context levels. In this dataset, a classifier trained on data from one country should be able to predict the stance of users in another. The different cultural and ideological motivations for the protests provide relevant roadblocks to test the robustness of the proposed estimators. However, given that these protests occurred concurrently, the dataset does not allow to effectively test this robustness over time. This is an outstanding problem in social media settings, as relevant textual features and prevalent memes can change significantly in the short-term. To address this issue, we constructed an additional weakly-labeled dataset around the 2020 Chilean Referendum.
The contributions of this chapter are the following:

- We propose a large-scale stance-detection architecture using transformers and graph neural networks that leverages a user's social network, a user's tweet timelines, and weakly-labeled protest-related tweets in order to predict their stance towards the government of each country. We also make publicly available a variant of a BERT language model [32], we call *twBETO*, developed for this work and trained on a substantial corpus of 150 million Spanish tweets. This model is not only trained on more tweets that other Spanish BERT models for Twitter[62], but has a better coverage of non-European Spanish dialects.
- We evaluate the effects of context on the performance of our classifiers in both in-country and related country-context settings. Through ablation studies, we examine cross-country performance of each stance model in a one-shot prediction setting.
- We collect a novel weakly-labeled Chilean referendum dataset to explore the ability of each Chilean model, trained at different context levels, to generalize to future data.

To the best of our knowledge, the evaluation of the effects of context on the generalization capabilities of these models has not been explored in other work, and has been identified as an outstanding issue in the task of stance classification [75]. Moreover, even though the subject of interest of our study are political stances in South American politics during charged political events, the method we have proposed can nevertheless be used in understanding social change and political movements in different cultures and continents.

4.2 Related Work

Stance detection is an essential component of many tasks associated with online social network moderation and analysis, including opinion polling and the detection of propaganda, misinformation, and hate-speech. In some domains, sentiment analysis may be a reasonable approximation of stance, but it has been shown that, on Twitter, sentiment polarity is a poor proxy, particularly around contentious political issues [92]. Tweets with a positive sentiment score can be used to oppose a topic and vice versa. As such, the overall sentiment of the text is not necessarily relevant to the classification [36]. Work in this area has concentrated in exploring stance in conversations (also known as rumor stance classification [141]) and on debates with respect to a predefined topic or target (known as target-stance classification). However, progress in this area has been limited by its reliance on small hand-labeled datasets, primarily created around challenge competitions like SemEval-2016 [92]. The work undertaken in this paper is an instance of

target-stance classification. In what remains of this section we provide a brief overview of the main research that have been applied in this area. For a more detailed description of these areas, and main datasets available, we refer the reader to Küçük and Can [75].

Target Stance Classification Target Stance Classification focuses on classifying the stance of a user or document with respect to a predefined topic or target. Task 6 of Semeval 2016 is a common benchmark for target stance classification. Authors have achieved SOTA results on this benchmark using architectures ranging from end-to-end neural ensemble models [108] to hand-crafted feature-based classifiers [3]. Due to the limited amount of data available, algorithms which rely on hand-crafted features are still prominent and achieve competitive performances with Deep Learning algorithms.

GNNs for Stance Detection Graph Neural Networks (GNNs) have become increasingly popular in recent years, but to our knowledge, they have not yet been used for stance-detection at the user level. Graph Neural Networks have been used in tasks related to stance in conversations at a document level, including several papers that explore GNNs for fake news classification and rumor detection. [133] propose a gated graph neural network model, PCGNN, for rumor detection on Tweet threads that were encoded with doc2vec. The authors achieve SoTA results on the PHEME dataset [72]. To our knowledge, all papers so far that have used graph neural networks for stance detection on Twitter have used individual tweets as nodes or graphs. This is effective for fine-grained stance detection, but no strong inference can be made about the alignment or intentions of the user that posted the false or misleading content. To combat misinformation at scale in an environment with state-backed propaganda campaigns, troll-farms, and bot networks, stance-detection at the user-level is essential for moderation at scale.

Context-Sensitive Stance Classification Context-sensitive classification, which models context in the form of spacial and temporal locality, is crucial to many tasks, including search engines, and context-aware Web applications [**?**]. This is highly relevant for stance detection in online social media where users engage with one another in a variety of ways, and thus focusing on an isolated sentence will likely result in an incomplete assessment of a user's stance [37]. Although some work have implicitly demonstrated the benefits of this approach by incorporating different levels of context in their architecture [**? ?**], there is a need to systematically quantify the resulting gains in classifier performance [75]. Importantly, to the best of our knowledge, no research has tackled the effect that leveraging the different levels of context available, has on the generalization capabilities of the proposed models. We aim to fill these gaps.

4.3 Data

We ground our analysis in the wave of protests that effectively paralyzed the South American region during the final months of 2019. We use the dataset constructed in Chapter 3, which identified the stances of thousands of Twitter users towards their respective governments during the event. The labels were constructed using a weak-labeling methodology that leveraged the users' endorsement of hand-labeled political figures as well as hashtag campaigns with well-defined stances towards each government. As shown in the previous chapter, these labels partition the users in communities that are polarized in their language and news sharing behavior. The protest dataset, which was collected between September 25 and December 24 of 2019, contains 550k labeled users split unevenly across the four countries and over 36 million labeled tweets. It contains an additional 1.1 million unlabeled neighbors and an additional 40 million unlabeled tweets. Stances are imbalanced in 3 of the 4 countries. In Table 4.1, we present the distribution of the different users in each country and the number of valid Spanish tweets available[1].

	Users			**Tweets**		
	Against	Pro	Neighbors	Against	Pro	Neighbors
Bolivia	58,508	54,347	292,684	3,508,300	3,583,943	12,507,155
Chile	220,391	33,331	409,014	14,659,535	2,775,496	14,211,964
Colombia	79,874	28,322	257,912	5,328,651	1,983,161	7,501,917
Ecuador	51,466	25,567	170,780	3,352,028	1,336,546	5,963,914

Table 4.1: Distribution of labeled users, their first-order neighbors (based on the response network) and their tweets (including retweets) for each of the countries studied.

For training purposes, we constructed training sets for each country by sampling 80% of the users (with their corresponding tweets), using 10% for validation and a 10% held-out test set [2]. Even though the regional nature of the dataset allows us to test robustness of our models at different context levels, it does not allow to effectively test this robustness over time. To address this issue, we constructed an additional weakly-labeled dataset around the 2020 Chilean Referendum.

The 2020 Chilean Plebiscite Throughout the 2019 Chilean protests, different social movements made calls in favor of drafting a new constitution. The social pressure exerted by protesters during the 2019 Chilean protests pressured the Chilean national congress to hold a National Plebiscite in 2020. It was scheduled for the start of 2020, but was delayed because of the Coronavirus pandemic. In October of 2020, it was overwhelmingly approved with 78% of the vote.

[1]Valid tweets included Tweets in Spanish, as determined by Twitter's API, with more than 5 tokens after the pre-processing step. This included the removal of all trailing hashtags, as the weak-labels were assigned in part by leveraging the usage of labeled hashtags at the end of a tweet.

[2]Note to reviewers: we plan to release both our code and the ids of the users used for training, validation and testing, but to uphold the anonymization policies for submission, we will make the link publicly available before publication.

	Against	Pro	Inconsistent
Tweets	10,084	9,626	1,117
Description	2,274	5,035	-
Both	2,409	1,476	6

Table 4.2: Final weak-label distribution of users based on the source used to assign the stance.

Using Twitter's v2 full-archive search endpoint feature available on their Academic Research Track[3], we collected tweets from September 25 to November 10 of 2020 (a month prior and two weeks after the plebiscite took place). We collected tweets matching 124 hashtags and terms relevant to the event (e.g.: *#Plebiscito2020*, *#PlebiscitoChile*, *#NuevaConstitucion*, etc.). We then labeled these hashtags to identify useful "Stance Tags"[4] and assigned weak labels denoting whether a user supported or opposed drafting a new Constitution.

The labeling procedure was as follows: an expert in the South American region who is fluent in Spanish reviewed a sample of tweets at the end of the collection period. The annotator then established if the tweets were used consistently in favor or against the approval of the new Constitution (or of the referendum process in general), or if this was not possible the hashtag was labeled "Undetermined". This was done in separate meetings where the reasoning for each label was openly discussed. This process resulted in 27 "Undetermined", 64 "Against", and 32 "Aprove" hashtags. Instead of using hand-labeled Political Figures as in Chapter 3, we opted to identify the hashtags in the user-description that were explicitly rejecting or approving the referendum. This resulted in 16 "Aprove" (*#Apruebo*, *#ApruboCC*, etc.) and 21 that "Against" (*#YoRechazo*, *#Rechazo*, etc.). This set is used to label users based on their user descriptions.

Following the methodology established in Chapter 3, tweets or descriptions were assigned a stance if they used stance tags with consistent stances. We only proceeded with users that had at least 8 tweets with a consistent stance or if at least one description was consistent. A stance was assigned to a user if at least 90% of their tweets had the same stance. For description-based stances, a user was assigned a stance if all different user descriptions observed during the period were assigned the same stance. The final user stance was determined by combining the labels obtained from both sources and validating the stance assigned to a subset of the users. The final number of user counts and their distributions are presented in Table 4.2.

Table 4.3 provides descriptive statistics of the labeled referendum users. We note that 45.5% of users labeled are not included in the 2019 Chilean Protest data. Moreover, to improve the resolution of the networks available for the labeled users, we collected their timeline during the event. Timelines were not always collected in the protest data, so we hypothesize that the additional timeline context for each user will improve the quality of user embeddings for the referendum data.

[3]This allows full historical access to publicly available tweets matching complex queries.

[4]Labeled Hashtags that are consistentlyy associated with a particular stance and appear at the end of a given tweet.

	Against	Pro	Neighbors
Users	10,423	10,423	96,202
Tweets	6,454,647	6,060,679	1,288,351

Table 4.3: Distribution of labeled users and the count of valid Spanish tweets (including retweets) by their stance towards the 2020 Chilean Constitutional Referendum. We also include counts corresponding to their first-order neighbors (based on the response network).

4.4 Methods

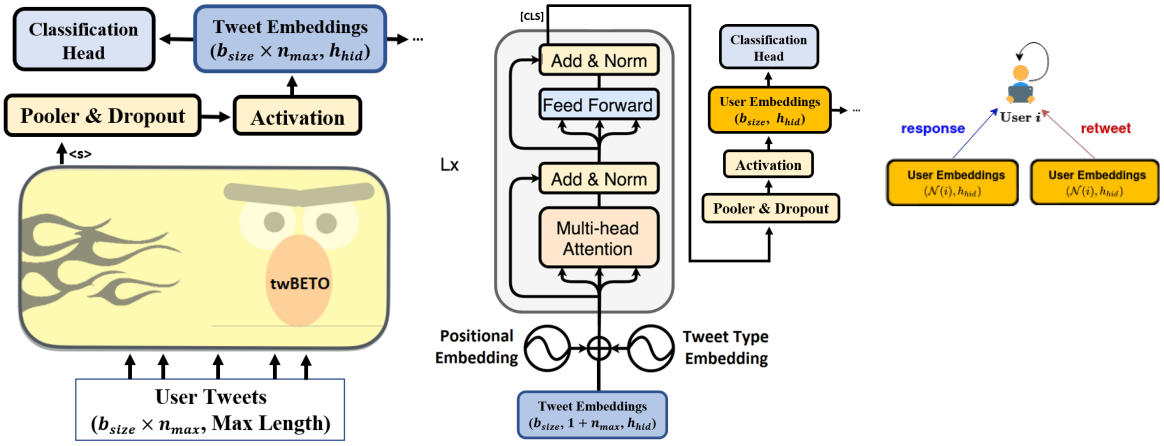

Figure 4.1: Proposed Architecture for the User Level Stance Classification. The left figure presents the Tweet Level encoding applied, which serves as input for the Transformer in the middle. b_{size} is batch size, n_{max} is the maximum number of tweets per users, Lx is the number of encoder stacks (3), and h_{hid} is the number of hidden dimensions. In the user level model, the model weights are back-propagated fully. On the right, we show the heterogeneous GAT model, wherein a user receives information from its neighborhood $\mathcal{N}(i)$.

The focus of this work is to explore the effect of increasing the level of context available to a classifier in the task of target-stance classification. As such, we made an effort to build a compartmentalized architecture which allows the usage of the output of each component to evaluate the baselines proposed. In Figure 4.1, we present the different context-based components of the architecture developed. Note that the output of each compartment not only serves as input for the next, but can be used for stance prediction with the information available at its context level. Training and predictions for each compartment was done independently, optimizing the cross-entropy loss using Adam with weight decay , a linear schedule with warmup and a maximum learning rate of 2e-4 (1e-3 for the network component). We train all models over 10 epochs with early stopping based on validation loss. Due to the unbalanced nature of the dataset in most countries (e.g.: Chile's label distribution was 87-13%), we opted for a dynamic resampling approach (with replacement) that weighted tweets and users based on the inverse frequency of their corresponding label. This under-samples the majority label and over-samples the minority label,

and is performed at the start of each epoch. We found that this strategy improved training stability and the validation performance of our transformer models when compared to other static over and under-sampling strategies. In what follows, we describe the different compartments of our proposed architecture. It is worth noting that only the Bolivian dataset was used to tune hyperparameters and select architectures.

4.4.1 Tweet Encoder

The first component of the model creates embeddings for the different tweets produced by the users. For this purpose we trained a BERT [32] language model, we call *TwBETO*, following the robust approach introduced in RoBERTa [86]. We opted for the smaller architecture dimensions introduced in DistilBERT [105], namely, 6 hidden layers with 12 attention heads. We also reduce the model's maximum sequence length to 128 tokens, following another BERT instantiation trained on English Twitter data [94]. We utilize the RoBERTa implementation in the Hugging Face library [129] and optimize the model using Adam with weight decay [69], a linear schedule with warmup and a maximum learning rate of 2e-4. We use a global batch size (via gradient accumulation) of 5k across 4 Titan XP GPUs (12 GB RAM each) and trained the model for 650 hours.
The model was trained with a corpus is comprised of 155M Spanish tweets (4.5B words tokens), as determined by Twitters API, and includes only original tweets (retweets are filtered out) with more than 6 tokens, compiled from the following sources:

- 110M Tweets (3B word tokens) from the South American protests collected from September 20 to December 31 of 2019.
- 25M (0.7B word tokens) Tweets collected around the Coronavirus pandemic from April 01 to December 31 of 2020.
- 3M (0.3B word tokens) Tweets collected around the Chilean referendum from September 25 to November 10 of 2020.
- 17M (0.5B word tokens) rehydrated targets across all the collections listed.

Tweets are pretokenized using the "TweetTokenizer" from the NLTK toolkit [18] and use the emoji package to translate emotion icons into word tokens (in Spanish). We also preprocess the Tweets by replacing user mentions with "*USER_AT*" and using the tweet JSON we replace media urls with "*HTTPMEDIA*" and web urls with "*HTTPURL*". We filter out retweeted Tweets, only keep tweets with more than 6 word tokens and truncate long tweets to 64 word tokens. The pretrained language model is made available in Hugging Face model hub.
The output of our *TwBETO* model, after being fed a batch of tweets, is pooled and passed through an activation layer. We use the standard pooling method used for BERT models, namely using the first element of the output of the final layer (corresponding to the "$\langle s \rangle$" token), as input for a fully connected layer and passed through an activation. This output serves as our final *Tweet Embeddings*. For the Tweet-level Context prediction, the embeddings are used as input for a classification head which is comprised of a connected layer and a softmax to predict the stance of a batch of tweets. Given that our main focus is to evaluate the effect of increasing the context

available to the model, we do not fine-tune the *TwBETO* parameters in any of the training settings.

4.4.2 User Encoder

As shown in the center part of Figure 4.1, the second component of our architecture is comprised of a stack of Transformer Encoder blocks [122] which operate on the *Tweet Embeddings* for a given user. As this requires a fixed input size, we introduced a parameter N that determines the maximum number of tweets to consider for each user. In this work we use $N = 15$. When users exceeded this limit, we sampled N tweets before assigning them to a batch. In this way we avoided wasting information as different tweets can be included each time a user is sampled[5]. In order for the *Tweet Embeddings* to serve as input for this component we reshape them and append a $[CLS]$ parameter vector at the start of each user's tweets, obtaining a tensor with dimensions: $b_{size}, 1 + n_{max}, h_{hid}$. Where b_{size} is the number of users included in the batch and $n_{max} \leq N$ is the maximum number of tweets observed for a user in the batch. We used trainable positional embeddings to maintain the temporality of the tweets and introduce a second type of trained embeddings to encode the type of each tweet as Original, Reply or Quote. Twitter users can interact in a variety of ways and we hypothesized that different interaction types would have different impacts on stance. This was confirmed in our ablation study (done for Bolivia), as the inclusion of each of these components improved the validation macro-F1 score at this context level. These 3 embedding tensors are then added and normalized (by layer normalization) and serve as input for our encoder stack. We use an encoder stack of size $L_x = 3$, based on the validation results.
The output of the last encoder layer is pooled following the same strategy used for the *TwBETO* model as this allows the $[CLS]$ parameter to attend to all tweets in the sequence and be optimized for stance classification. The output of this pooling layer serves as our final *User Embeddings*. As before we feed the embeddings to a classification head to perform user-level classification, or use this as input for the next component or our architecture.

4.4.3 Network-Based Prediction

We take the embeddings produced by the user encoder and predict users' stance using their interactions in the social network. We explored several different graph neural network architectures, including GraphSage layers, Graph Convolutional layers, and W-L Graph Convolutional layers. Ultimately, we achieved the highest accuracy using Graph Attention Network (GAT) layers proposed by Veličković et al. [123]. In the social setting imposed by Twitter's platform, there are also different context levels that can be leveraged by our Network-based classifier to improve stance detection. To test this, we trained separate homogeneous models for the network retweets, the network of responses (the union of replies and quotes) and the combined network (obtained by the union of both). Finally, we trained a heterogeneous model where we distinguish response and retweet edges.

[5]Given the power law distribution of user tweet counts and to reduce sampling times when constructing the batches, we chose to only keep the first 150 tweets available for each user.

For homogeneous graph experiments, after testing different alternatives, we use two GAT Layers with dropout, each with four attention heads, followed by 4 batch-normed linear layers with dropout and relu activations. The corresponding output is fed to a classification head for our final network-based user stance classification. We observed our heterogeneous model quickly overfit the data with our homogeneous architecture, so for the heterogeneous model, we removed 2 linear layers and use only one attention head (for each edge type). Given the size of the underlying graphs, we sample neighborhoods for each node using the neighbor loader proposed in Hamilton et al. [55]. In our GAT models, for a given user i, the input to the attention layer are the *User Embeddings* of a random sample of k one-hop neighbors of i denoted as $\mathcal{N}_k(i)$. In our homogeneous GAT networks, the attention mechanism calculates the importance of $\mathcal{N}_k(i)$ to the label of i. In our heterogeneous GAT networks, the attention mechanism calculates the $\mathcal{N}_k(i)$ with relation Φ (retweet or response) to the label of i.

4.5 Results and Discussion

4.5.1 Main Results

In this section we present the results of the different components of our architecture to evaluate the effects of increasing the level of context available to a classifier. We use the Tweet-level classification as a baseline for our ablation studies. It is important to note that this task is evaluated at a tweet level (not at a user level) by assigning each its corresponding user stance. In this way we are able to define four different experiments to assess the performance of classifiers trained at this context level.

- *Weak-Labeled Tweets*: This Tweet-Level context exercise is trained using tweets that contained "Stance Tags" as identified in [125]. To avoid overfitting to these tokens, one of our pre-processing steps removed all trailing hashtags.
- *All Original Tweets*: This Tweet-Level context exercise is trained using all original, replies and quote tweets produced by the different weakly-labeled users.
- *User Average of Original Tweets*: This User-Level prediction exercise is done by averaging the predicted labels assigned to a given user. A user is assigned the majority label of its tweets and when none exists a stance is assigned randomly.
- *User Average of All Tweets*: This User-Level prediction exercise is similar to the one described above, but also includes the retweets done by a user. Retweets are assigned the label predicted for its target.

We include the last two user-level predictions to evaluate how effective our proposed User Transformer is at aggregating tweet information. We evaluate the performance of each model using accuracy and macro F1 (due to the observed asymmetries in the label distribution) on the held-out test set. In Table 4.4, we present the results of the ablation studies. As shown, there is a clear improvement in performance when more context is available to the classifier. As expected, focusing only on the weakly-labeled Tweets presents an easier task than focusing on all original tweets, but still falls short to User-Level averages. When retweets are included in the averages, the performance improves, but still lags behind the User Transformer in all countries. Interest-

ingly, we observed that for Colombia, adding retweets hurts the user average performance. The User Transformer does not exhibit these disparate behaviors, highlighting its ability to identify tweets that are more relevant to the stance of the user. Importantly, and as hypothesised at the start of this work, leveraging the social network of the user can consistently improve upon the already high bar of the User level Transformer.

			Bolivia (%)		**Chile (%)**		**Colombia (%)**		**Ecuador (%)**	
			Acc.	M-F1	Acc.	M-F1	Acc.	M-F1	Acc.	M-F1
Tweet-Level	Weak-Labeled Tweets		85.48	83.90	83.75	73.82	77.42	76.53	82.47	82.14
	All Original Tweets		75.50	75.49	72.03	66.83	69.95	65.55	73.45	71.12
User-Level	Average Tweet-Level	All Original	79.92	79.62	86.06	76.26	81.04	73.87	82.24	79.17
		With Retweets	89.09	89.03	94.83	88.78	81.49	67.86	92.35	90.70
	Transformer		94.05	94.05	95.98	91.53	95.62	94.26	94.99	94.34
Network-Level	Homogeneous	Response	94.02	93.94	95.60	90.56	95.13	93.60	94.19	92.88
		Retweet	94.00	94.00	96.23	91.53	95.69	94.40	95.03	94.38
		Combined Network	94.06	94.06	95.74	90.00	95.61	94.25	94.65	94.01
	Heterogeneous		**95.77**	**95.77**	**97.35**	**94.29**	**96.92**	**96.01**	**96.23**	**95.77**

Table 4.4: Performance of in-country Stance Classifiers at different context levels.

4.5.2 Robustness Analysis

Cross-Country Robustness Even though the addition of social context can lead to an in-sample increase of classification performance, we wanted to test how context can affect the generalizability of our proposed classifiers. Given that our target countries share a common language, we hypothesized that our models should be able to extrapolate stance learned in country to another. Protests in Bolivia had opposing ideological motivations than the ones observed in the other countries, which we hypothesized would degrade performance significantly. The results for the cross-country ablation are shown in Table 4.5. For economy of space, we only include the macro F1 score for the best performing classifiers at each context level (for the Tweet-Level scores we use the User Average of tweet predictions including retweets).

As shown, the Bolivian case serves as an adversarial setting for classifiers trained in other countries, which suggests that, when applied to this country, the semantic features leveraged by the classifiers are operating on an ideological dimension. This is consistent with results presented in Chapter 3, where we showed that language polarization remained along ideological lines when comparing protests of any of the three countries with Bolivia. This was not the case when comparing protests of the other three countries. As expected, Chile, Colombia, and Ecuador exhibit strong pairwise performance consistent with their ideological alignment. We can also note that the Tweet-Level performance varies significantly, falling in some cases 30 points below the user Transformer, which is far more stable (75-85% in these countries). We found that on average, excluding Bolivian predictions, the heterogeneous network model increased cross-country prediction macro F1 score by 10.1 points (in the best case the Ecuadorian instance yielded a zero-shot macro F1 of 95.5 on Colombian data). It is worth noting that although the heterogeneous model achieves this on countries with similarly motivated protests, it also significantly under-

performs other alternatives in the Bolivian case. This is consistent with the observed tendency of deep learning models to be overconfident in their predictions, and highlights the importance of domain knowledge when applying these models to different domains. If we would leverage the knowledge of the diametrically opposed motivations for the Bolivian protests, by inverting the Heterogeneous model's predictions, then its performance would be in line with what was observed for the other countries. We take this approach next, when we explore the predictions for the 2020 Chilean referendum.

		Bolivia	**Chile**	**Colombia**	**Ecuador**
Bolivia	Av. Tweet-Level (WR)		10.39	17.66	22.79
	User-Level Transformer		9.95	17.36	21.48
	Hetero. Network-Level		3.84	5.78	22.36
Chile	Av. Tweet-Level (WR)	31.83		90.01	62.17
	User-Level Transformer	30.88		84.86	83.40
	Hetero. Network-Level	21.54		90.85	88.91
Colombia	Av. Tweet-Level (WR)	36.56	81.98		48.41
	User-Level	35.32	75.22		78.89
	Hetero. Network-Level	31.24	86.51		90.55
Ecuador	Av. Tweet-Level (WR)	31.01	57.99	78.98	
	User-Level Transformer	21.18	80.92	78.32	
	Hetero. Network-Level	10.67	89.78	95.50	

Table 4.5: Macro F-1 (%) score for out of sample cross-country predictions for classifiers at different context levels.

Robustness over time To test the effect of context over time, we applied the Chilean classifiers to predict the stance of users, in a zero-shot setting, towards the 2020 Plebiscite vote for drafting a new constitution. We hypothesized that opposition to the government during the protests should signal endorsement for the new constitution (the models should be good inverse classifiers). Table 4.3 presents the results of this exercise. As extra validation, we also include the predicted Referendum vote based on the two-hop neighborhood of labeled users. As shown, the results are consistent with what was described before, with more context improving the robustness of a classifier. We note that the heterogeneous model improves its protest performance considerably in this setting. We hypothesize that this is due to the better resolution of networks constructed for the labeled users (as we collected their timelines during the event). With regards to the predicted referendum vote, we note that semantic-based classifiers tend to undershoot the observed Referendum tallies (the final vote was 78% in favor of the new constitution), while the single-network based classifiers do the opposite. Interestingly, the heterogeneous model predicts the vote almost perfectly which is encouraging, albeit more research is necessary to assess if this behavior is generalizable in other countries. The network models' strong performance can be in

part explained by the fact that language and conversation topics change faster than social ties, and the fact that social ties themselves can alter language [77]. As a result, features extracted from interaction networks may carry strong signals that are more generalizable to different cultural contexts.

			Acc.	M-F1	Pr-Ref (%)
Tweet-Level	Weak-Labeled Tweets		70.19	72.03	57.82
	All Original Tweets		67.75	68.06	65.18
User-Level	Avg. Tweet-Level	All Original	82.60	83.18	63.81
		With Retweets	84.33	86.14	67.81
	Transformer		90.21	90.22	63.36
Network-Level	Homogeneous	Response	81.43	84.18	86.3
		Retweet	94.07	94.31	82.3
		Combined	77.05	81.29	87.6
	Heterogeneous		**99.74**	**99.74**	**78.5**

Table 4.6: Out of sample Predictions for the Chilean Referendum at different context levels. These correspond to the classifier trained on the 2019 Chilean Protest Data, but with inverted labels ("Pro" government are considered "Against" the referendum an vice-versa).

4.6 Discussion

For each country we examined, our heterogeneous graph neural network yielded the highest accuracy and F1 scores. In all four countries, the heterogeneous model yielded macro F1 scores greater than 94.2 and accuracy scores greater than 95.7. The heterogeneous network model yielded an average 1.9 point increase in macro F1 score over the transformer model alone.
The strong performance of the heterogeneous model aligns with intuition, as different edges on Twitter have different social functions. Retweeting with no commentary is more likely an endorsement than an argument, whereas replying could imply a fight or it could imply an agreement. To test this explanation, for each node i in labeled Bolivian retweet and response networks, we calculated the percent of neighbors of node i that shared the same label as node i and then took the average of those percentages. In the retweet network, the average of neighborhood-agreement percentages was 93%. In the Response (reply, quote, mention) networks, neighborhood-agreement was only 69%. This suggests that ability to differentiate between relation type should help the model's performance. Our results confirm that including interaction types in models can yield modest improvements in predictive power in in-country stance detection tasks. However, we find that the inclusion of network data results in yields much larger improvements in related country-context assessments.

4.7 Limitations and Future Work

We use binary stance labels in our models, so users with no opinions or contradictory/nuanced opinions may be incorrectly categorized into pro- or anti-categories. Finally, we did not attempt to remove bots or trolls from the datasets, nor did we survey in-country Twitter user demographics, so these datasets may be noisy proxies for population opinions. However, despite not removing bots and trolls, we found the stance distribution of users in the Chilean referendum data was nearly identical to the final referendum vote.
Additionally, our network model is only using interactions (reply and quote, retweets) and the text of these users for classification. This excludes many other attributes available on Twitter that might correlate with stance, including URLs, followership, posting patterns, bios, and shared multimedia. In future work, we plan to explore integration of multimodal data as well as the integration of other attributes that may be useful in detecting stance.

4.8 Conclusions

In chapter we explored the value of context in the task of target-stance classification during the 2019 South American Protests. For this purpose we constructed a compartmentalized architecture that relied on Transformers for the Tweet and User level contexts, and GNNs to leverage social media relations. We found that increasing context not only improved the performance of a classifier within the country it was trained, but also made it more robust to out-of-sample predictions. We found these out-of-sample improvements were substantial both when comparing a classifier's performance across varying country contexts and over time.

Chapter 5

Stance in Replies and Quotes (SRQ): A New Dataset For Learning Stance in Spanish Twitter Conversations

5.1 Introduction

In recent years, the prevalence of fake news on online social networks and their perceived increasing levels of polarization have sparked the interest of the international community. In academia, much of the effort has been spent in the automatic identification of false or rumorous information [27] and the detection of controversy as an indicator of polarized discussions [46]. Stance detection in conversations, classifying a response according to whether it agrees (or supports), disagrees (or denies), or comments its target, is increasingly gaining popularity as an auxiliary subtask for both these areas. However, despite the importance of this subtask in identifying rumours [39, 71] and controversy [2, 75], the corresponding methods, and datasets they rely on, have largely been developed independently.

Consequently, we can identify three significant limitations in the existing resources available for the detection of stance in Twitter conversations: 1) The majority of annotated datasets are built around rumorous tweets in order to determine the veracity of said posts based on stance taken in their replies [52, 141]. Though useful for rumor detection, this does not generalize to non-rumor events [22] and limits their usefulness for assessing polarization or controversy. 2) The vast majority of existing datasets focus primarily in direct responses and do not take into account quotes. This is critical as quotes have been gaining prominence since their introduction by Twitter in 2015, especially in the context of political debates [45] or as way of debunking disinformation [11]. 3) Even though some non-English datasets for the detection of stance towards a predefined target have been developed [117, 136], to the best of our knowledge, none exists for the detection of stance in conversations. This has limited the exploration of the effects of rumors and polarization on online discourse in non-English speaking regions.

In this work, we construct a labeled dataset to begin to address these limitations. We focus on a major political event that shocked the South American Region at the end of 2019, and whose consequences still shape the current political landscape [130]. Namely, we consider Twitter con-

versations (in Spanish) around the protest and subsequent riots that started in Ecuador, followed by Chile, Bolivia and Colombia at the end of 2019. These events not only paralyzed these countries for weeks, and in some cases months, but also had a massive online presence. As we show in Chapter 3, there are also several reported instances of disinformation campaigns that pervaded the online discussion of these events. We adapted a previous sampling methodology to sample contentious conversations [76], and annotated a sample of 7.4 thousand target-response pairs from the universe of collected tweets. It is worth noting that our sampling methodology stratifies responses on whether they are Replies or Quotes to ensure adequate sample representation for each type.

Importantly, we propose a unified framework for the annotation of data for the detection of stance in conversations. Recent work on stance labeling in social media conversations has centered on identifying 4 different positions in responses: support, denial, comment, and queries for extra information [101]. This limits the annotation of positive or negative instances, to interactions that either support or deny the veracity of the target statement (or the associated rumor). This prevents the identification of more general forms of agreement/disagreement between users and, as mentioned, limits its usefulness for assessing polarized discussions. For this reason, we separate the task of detecting stance in conversations in two parts: first we identify whether the response agrees, disagrees, comments, or queries its target, and then we determine whether the agreement (disagreement) is actually supporting (denying) the veracity of said statement. In addition, we also identify the stance of the target and response tweets towards the protests and government in each country (a target-stance detection task). We believe that this unified treatment can help bridge these two related research areas and is necessary in order to explore how polarization in online discussions can affect the spread of rumors.

Finally, we train a second version of the *twBETO* language model, this time on 200 million Spanish tweets (50 million more than before), with a better distribution of Spanish variants and with an improved tokenization method. This new language model serves as the base component of a classifier specialized on Spanish Twitter conversational data and is trained on the joint Conversation Stance classification task previously described. In this setting, we are able to achieve improvements on the prediction's macro F1-score, when compared with existing English datasets for this task.

5.2 Related Work

Supervised text classification models with deep learning techniques are able to estimate a document's class with highly accuracy [38]. However high-quality training examples are required, which generally entails active human participation to mitigate concerns, increase inter-annotator reliability among annotators, and lower topic drifting when applied to different domains. Hence, one of the main barriers to overcome in this area is the lack of annotated resources, especially in Spanish. Machine learning techniques, such as: Linear Discriminant Analysis (LDA), Random Forest, or Support Vector Machines (SVM) have also been used as classifiers to uncover word patterns in predicting an output about elections [59]. The most successful current architectures for Natural Language Processing (NLP) tasks, such as BERT [32], allow learning bidirectional contextual word representations, where words are represented by different embed-

dings based on the context of the word. This makes it possible to learn complex structures and word meanings, such as polysemy or homonym [32]. However, it has been noted that that the models' performance deteriorates as the target domain moves away from the pre-trained domain [32, 60, 87, 137]. The ability of Transformer-based architectures to learn new tasks from explanations in large language models has also been explored [6, 62, 140], where authors found that inter-sentence coherence played a key role in the model's performance, in particular when the task is about reasoning on pairs of tweets. With regards to the sub-field of learning stance from data, the application of Transformer-based architectures has been limited due to the size of available annotated datasets, while feature-based alternatives have remain strong competitors.
Topics on learning stance from data could be broadly categorized as having to do with: 1) Stance in posts on social media, and 2) Stance in Online Debates and Conversations. The recent boost in research on both forms of stance detection in social media has been driven by annotated data made available for competitions like SemEval (for English tweets), NLPCC-ICCPOL (for Chinese microblogs) or IberEval (for Spanish tweets). In what follows, we provide a brief overview of these topics. For a more detailed description of the current resources available for this task, refer to the survey by Küçük and Can [75].

Stance in Social-Media Posts One of the earliest competitions in the subject was Task #6 of SemEval 2016, where Mohammad et al. [92] built a stance dataset towards predefined targets, using English tweets of several different topics. Many researchers [8, 85, 127] used this dataset and proposed algorithms to learn stance from data. However none of them exceeded the performance achieved by a simple algorithm [92] provided as a baseline. The success of the event spurred similar competitions in other languages. In one of the few Spanish examples of this kind, Taulé et al. [117] developed a similar type of annotated dataset around the Catalan Independence protests for the IberEval 2017 conference. Given the small-scale of the datasets available, feature-based classifiers show a strong performance in this task, often outperforming deep learning alternatives [92, 117].

Stance in Online Debates and Conversations The idea of stance in conversations is very general and its research origin can be traced back to identifying stance in online debates [111]. Though stance-taking by users on social-media, especially on controversial topics, often mimic a debate, social-media posts are very short. An approach of stance mining that predicts stance in replies – categorized as 'supporting', 'denying', 'commenting' and 'querying' – to a social media post is gaining popularity [141, 143]. Prior work has confirmed that replies to a 'false' (misleading) rumor are likely to have replies that deny the claim made in the source post [144]. Therefore, as was discussed in Chapter 2, this approach is promising for misinformation identification. However, the earlier stance datasets on conversations were collected around rumor posts [141], they contain only replies, and have relatively few denials. Our new dataset generalizes this approach and extends it to quote-based interactions on controversial topics. As described, this new dataset is distinct as: 1) it distinguishes between 'replies' and 'quotes', the two very different types of interaction on Twitter, 2) it is collected in way to get more non-neutral stance examples, which were a minority label in [143], and 3) it is collected on general controversial topics and not on rumor posts. Importantly, to the best of our knowledge, no resources for this

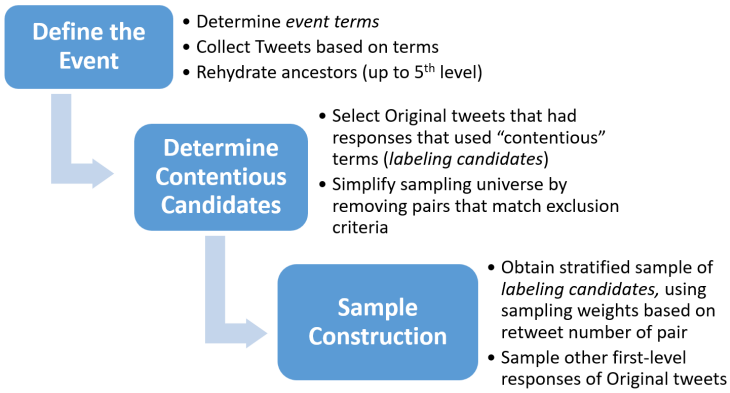

Figure 5.1: Methodology used for the construction of the sample to be annotated.

task are currently available in Spanish.

5.3 Methods

5.3.1 Dataset Collection Methodology

We adapted the methodology developed in [76] to the Spanish language and the context of the protests. This sampling scheme results in a more balanced fraction of denial responses, by skewing towards contentious or controversial interactions. Importantly, one limitation noted is that it identifies controversial source tweets by focusing on responses that contain "contentious" terms. Although this is an effective way of identifying these interactions, focusing on sampling only these responses may hinder a model's ability to generalize to other ways of expressing support or denial. To address this, we recover more of the identified contentious conversations by also sampling other direct responses not containing these terms. In Figure 5.1, we summarize the main steps followed to obtain the final sample which we describe in more detail next.

Step 1: Define the Event Our focus is on the Twitter conversations surrounding the protests that paralyzed the South American region at the end of 2019. For this purpose, we use the dataset described in Section 1.2, which is comprised of over 100 million tweets, from 15+ million users, collected using approximately 500 hashtags and terms (denoted *event terms*). As noted, an effort was made to maintain the conversational structure of the data by rehydrating ancestors (up to the fifth degree) that were not included in the collection and to collect data around antagonistic positions by including hashtags used more prominently by either supporters or opponents of the different protests.

Step 2: Determine Contentious Candidates A conversation thread is selected as potential candidate to be annotated if the source target contains any of the *event terms* and any of its responses contain a contentious term. We denote these origin tweets as *contentious candidates* and the responses containing the contentious terms as *labeling candidates*. The set of contentious

terms used includes 139 terms or phrases that could indicate that the replier accuses the target of lying or spreading disinformation. This includes terms like ”mintiendo” (lying), ”incorrecto” (wrong), ”engañoso” (deceitful), ”falso” (false), etc. The full list of terms used is provided in the appendix.

To reduce the sampling universe, we filtered the candidates based on some additional conditions. We only used tweets that were identified by Twitter to be in Spanish and excluded responses from a user to herself (as this are used to form tweet-threads). In order to simplify the labeling context, we also excluded responses that included videos, or that had targets that included videos and limited our sample set to the first level of the conversation tree (by including only direct responses to the *contentious candidates*). We show the distribution of the resulting sample universe in Table 5.1.

	Labeling Responses		**Other Responses**	
	Replies	Quotes	Replies	Quotes
Bolivia	350 222	96 339	1 401 047	3 083 485
Chile	443 317	140 390	5 019 874	2 678 292
Colombia	314 702	87 609	3 074 367	1 385 512
Ecuador	351 523	100 746	3 310 067	1 508 016

Table 5.1: Distribution of relevant tweet pairs by response type that define the sampling universe.

Step 3: Sample Construction The final sample was obtained by stratifying the universe of *labeling candidates* based on the following criteria: the country event (defined in the collection), the response type (Quote/Replies), and the “contentious” terms. The last two stratification levels were included to ensure adequate representation of the different response types and also sufficient coverage of the language used to signal contention. In addition, sampling weights corresponding to the logarithm of the total retweet number[1] of the target-response pair were used. We relied on this human signal of interest in the conversation as its necessary precondition for both rumours [142] and controversy [46].

To ensure that the final sample also includes other ways of denying/supporting a target tweet, an additional sample of other responses to the sampled *contentious candidates* (not including the contentious terms) was selected. The final distribution was 70% *labeling candidates*, 30% other responses. Table 5.2 presents the final distribution of the final sample comprising 7 395 annotated pairs. As shown in the table, the dataset released also includes the tweet IDS of all collected unlabeled responses (in Spanish) that were part, at any level, of the conversation tree of the labeled pairs. We expect that the rich network structure will be helpful to improve the performance of classifiers trained with this data.

[1]We utilize the logarithm to account for the power-law distribution of the retweet number and maintain randomness in the final sample.

	Quote	Reply	Original Tweets	Unlabeled Responses
Bolivia	1009	720	1144	551 574
Chile	804	763	1247	696 907
Colombia	951	817	1116	460 588
Ecuador	770	668	1018	193 810
Other	370	352	590	251 583
International	78	93	151	91 664
Total	3982	3413	5266	2 246 126

Table 5.2: Final Sample Distribution based on the country referenced in the tweet pair as identified after the annotation process. The International category comprises tweets that referred to more than one country.

5.3.2 Annotation of Sampled Conversations

The annotation process was handled by native Spanish speaking students of two Ecuadorian Universities. Before being selected as annotator, the students were shown an hour long presentation describing the project and asked to review a manual that described the events and provided annotated examples explaining the reasoning behind each choice (the Spanish manual is included in the supplementary information). The students interested in continuing with the process were then tested on the complete annotation of 5 target-response pairs that were previously annotated by the authors of this paper (labels were decided after reaching a consensus in a group discussion). We used an 80% cut-off as a requirement to continue, which resulted in 98 different annotators. Each target-response pair was labeled 3 times by different annotators and the final label decided by simple majority. If a majority was not reached for any of the main stance questions (see below), the pair was flagged as problematic and set aside to be resolved by arbiters. More details of the arbitration process are provided when describing the web platform developed for the annotation.

Annotation Task The annotation task is divided in the following 4 sections:

1. **Event identification:**
   (a) What country is being discussed in the original tweet?
   (b) Are the tweets related to the protests that occurred in the country(ies) identified in the previous question?
2. **Target's stance towards the protests**:
   (a) What is the stance of the Original tweet with respect to the protests?
   (b) What is the stance of the Original tweet with respect to the government?

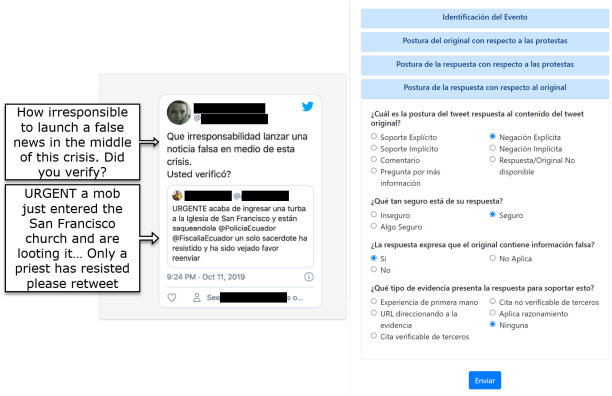

Figure 5.2: Rendered Form for the annotation of a Quote-type response to a *contentious candidate*. The English translations are added for illustrative purposes.

3. **Response's stance towards the protests**:
   (a) What is the stance of the Response tweet with respect to the protests?
   (b) What is the stance of the Response tweet with respect to the government?
4. **Stance of the response towards the target**:
   (a) What is the stance of the Response tweet towards the Original tweet?
   (b) Does the Response tweet express that the Original contains false information? (Conditional on 4.a)
   (c) Does the Response tweet express that the Original contains true information? (Conditional on 4.a)
   (d) What type of evidence is presented by the response to support this? (Conditional on 4.c or 4.d)
   (e) How certain are you of your answer?

In Figure 5.2 we provide an illustrative example of a Quote-type interaction, discussing a rumor during the Ecuadorian protests, and the corresponding annotation form as presented in the web platform. The purpose of the first section is to identify if the interaction refers to: a) Bolivia; b) Chile; c) Colombia, d) Ecuador; e) Others; f) Not clear (multiple answers are possible). In addition, it is necessary determine if the tweets are referencing the protests that occurred in the identified country. The second and third sections are target-stance classification tasks for the Response and Original tweets respectively. Here we need to infer the stance of both tweets towards the government and the protests that transpired and in both cases the possible categories are: a) In favor; b) Against; c) Neutral; d) Not clear; e) It does not apply. The final category was reserved for tweets that did not discuss the protests or the government. In the illustrative example presented, the Response tweet has a Neutral language to both the protests and the government, while the Original is Against the protests but Neutral to the government.
The final section encapsulates the conversation stance detection task. The annotator is first required to identify the stance of the response towards its target among 6 possibilities: a) Ex-

plicit Agreement; b) Implicit Agreement; c) Comment; d) Ask for additional information; e) Explicit Disagreement; f) Implicit Disagreement. We distinguish between "Explicit" and "Implicit" forms of (dis)agreement following [76], although in that work it was applied to the rumor stance detection task. The former type refers to responses that include terms that explicitly state that they (dis)agree with their target (e.g. 'That is a blatant lie!'). The implicit category on the other hand, as its name implies, corresponds to responses that do not explicitly mention the stance of the user, but that, given the context of the target, are understood as (dis)agreements. These are much harder to classify, as they can include sarcastic responses, and can be leveraged for error analysis of trained classifiers. They can also be helpful for the annotation process, as was confirmed by feedback given by the annotators. The second component of the conversation stance detection task is only required for responses that agreed or disagreed with their target (either explicitly or implicitly). Annotators were asked to determine whether the response accuses the target of lying or spreading disinformation (q. **4.b** shown if they disagreed) or if it affirms the target is telling the truth (q. **4.c** shown if they agreed). Whenever the question did not fill the requirements to be shown, the system assigned the "Does not apply" label. Finally, if the annotator answered positively to the last question assigned, they are queried for the level of evidentiality presented [141]: a) First-hand experience; b) URL pointing to the evidence; c) Non-verifiable quote from third party; d) Verifiable quote from third party; e) None. In the case of the example presented, and as shown in Figure 5.2, the response not only Explicitly Disagrees with the Original, but also accuses it of spreading disinformation while not providing any evidence for the claim.

Annotation Platform To facilitate the annotation of the *contentious candidate* pairs, and to accommodate to the context of the Coronavirus pandemic, it was necessary to develop a web application. Each annotator received credentials to access the application as one of two roles: *Annotator*, or *Arbiter* for conflict resolution. The later credentials were reserved for the authors of this work that are native Spanish speakers and paid research assistants that were an active part of the project. In total, there were 8 different arbiters. When annotating a tweet pair, the platform rendered an Original tweet and its corresponding response along with a form containing the questions described before (see Figure 5.2). The tweets were displayed by querying Twitter's API, so they were also valid hyperlinks to the corresponding conversation. The annotators were encouraged to follow these links if uncertain of their answers in order to gain additional context of the tweeters or the conversation.

The selection of tweets to be displayed was made randomly from the set of *contentious candidate* pairs that had not been assigned the three required annotations. A set of control filters was put in place to ensure *Annotators* were presented only tweet pairs they had not already annotated. After completing the form, a "Submit" button became available to send the answers to the server, which resulted in another random pair of tweets displayed to repeat the process.

The application was also designed for the resolution of tweets identified as "problematic", which were flagged automatically if, after being assigned 3 different annotations, a majority label was not reached (by simple majority) in any of the main questions[2]. When logging in on an *Arbiter*

[2]This main questions considered were: 1.b, 2.a, 2.b, 3.a, 3.b, 4.a, 4.c, 4.d. Refer to **Annotation Task** for more details in the questions.

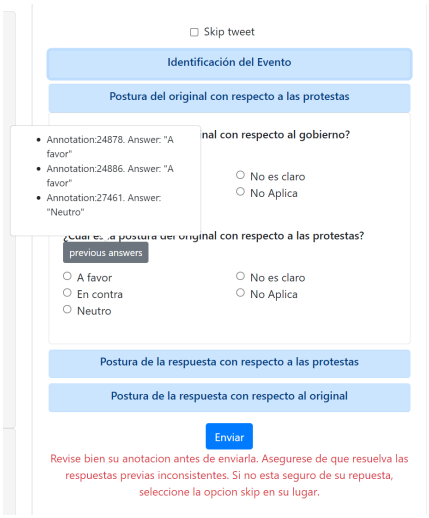

Figure 5.3: Form presented to *Arbiters* for the resolution of Problematic Tweets. The corresponding tweet pair is omitted for economy of space. This form includes extra functionalities necessary for the resolution task, like displaying previous annotations for each question and the ability to flag the pair as "Skipped" for extra deliberation.

role, the section "Problematic Tweets" became available, were the application displayed the list of all tweets to be resolved. The pairs to be resolved were rendered using the same interface as in the annotation step, but with some added functionalities. For a given rendered problematic pair and each of the questions, the *Arbiter* had the ability to query all the annotations available. The form also had the option to "Skip" a problematic pair which, when selected, internally flagged the pair as skipped for additional deliberation. *Arbiters* were instructed to skip a tweet when: they were not confident of the correct majority, their choice was not one of the options selected by the annotators, or when they disagreed with the majority determined in non-problematic questions. Problematic pairs flagged as skipped were moved to the end of the list, and were only resolved in a group setting via discussions of two or more *Arbiters*.

The application was developed in Python version 3.6.8, using the Django web framework version 2.2.4. The framework was chosen because it contains a robust Object-Relational-Mapper (ORM) to make the management of the data easier, and also comes with ready-to-use modules for authentication and authorization schemas. The Database used was MySQL version 0.8.21. And the application was deployed in a CentOS Linux server with an Intel(R) Xeon(R) Silver 4114 processor using 8 cores and 8 GB of RAM.

5.3.3 Proposed Classifier

***twBETO* v1** For this work, we trained an upgraded version of the model used in Chapter 4, using the same robust pretraining approach introduced in RoBERTa [86]. We maintained the same architecture as before, namely, 6 hidden layers with 12 attention heads and a maximum sequence length to 128 tokens. We utilize the RoBERTa implementation in the Hugging Face library [129] and optimize the model using Adam with weight decay [69], a linear schedule with

warmup and a maximum learning rate of 2e-4. We use two NVIDIA RTX A6000 (48 GB RAM each), a per-device batch size of 230 and accumulate gradients for 11 steps to obtain a global batch size of size 5060 and trained the model for 500 hours.
The model was trained with a corpus is comprised of 201M Spanish tweets (5.5B words tokens), as determined by Twitters API, and includes only original tweets (retweets are filtered out) with more than 6 tokens and truncate long tweets to 64 word tokens. The sources of the final data used to train the model are distributed as follows:

- 92.2M tweets from the South American protests collected from September 20 to December 31 of 2019.
- 56.7M tweets collected around the Coronavirus pandemic from April 01 to December 31 of 2021.
- 11.6M Tweets collected around the Chilean referendum from September 25 to November 10 of 2020.
- 9.1M tweets collected around the 2021 Colombian protests from April 15 to June 09 of 2021.
- 31.4M tweets collected by sampling contentious responses and rehydrating up to the fifth level ancestor in the conversation tree. The contentious term list contains 166 entries and expands on the one defined previously by including phrases for supporting the target. Data was rehydrated using Twitter's v2 full-archive search endpoint feature available on their Academic Research Track by sampling every third month of the year from January 2019 to December 2021.

Tweets are pretokenized using the "TweetTokenizer" from the NLTK toolkit [18] and use the emoji package to translate emotion icons into word tokens (in Spanish). We improve the tokenization method used in the model *v0* by preserving more information on the specialized tokens used to replace URLs and Mentions. Namely, mentions that are included as an artifact with a reply are replaced with "REPLY_AT", while normal mentions are replaced with "MENTION_AT". Similarly, for URLs we include different masks corresponding to the most popular domains: "MEDIAURL" (for urls corresponding to embedded pictures or videos), "TWTURL" (for Twitter), "YOUTURL" (for Youtube), "FACEURL" (for Facebook), "INSTAURL" (for Instagram), "URLdGOV" (for gov or gob top-level domains), "URLdRU" (for ru top-level domains), "URLdEDU" (for edu top-level domains), "URLdORG" (for org top-level domains), and "HTTURL" for any other URL. The pretrained language model is made available in Hugging Face model hub.
The output of our *TwBETO* model, after being fed a batch of tweets, is pooled and passed through an activation layer. We use the standard pooling method used for BERT models, namely using the first element of the output of the final layer (corresponding to the "$\langle s \rangle$" token), as input for a fully connected layer and passed through an activation and fed to a classification head. In addition, we introduce a novel auxiliary pretraining objective, we call Response Prediction Task (RPT), that relies on unlabeled Twitter interactions.

Response Prediction Task (RPT) The proposed auxiliary objective requires the model to determine whether an input pair of tweets are in a target-response relationship. For any Twitter collection around an event, especially ones that follow the standards proposed in section 1.2, we can construct a balanced dataset for classification by selecting a sample (or all) of target-response pairs available. The negative classes can be easily obtained by duplicating the targets, but this time assigning a random response to it. In this work, we ensure that the following criteria are met by any selected negative sample: 1) is not a response to the same tweet, 2) is not a response to any tweet in the conversation tree of the target, 3) The author of the response differs from the author of the target. This auxiliary task, reintroduces the next-sentence prediction objective removed by the RoBERTa methodology, but adapted to a Twitter setting. As we show in this work, this can improve the performance of the model for downstream tasks, especially in the context of small datasets.

5.4 Results

In this section we present the results of the annotation process and several baseline classifiers for the different questions.

5.4.1 Annotation Results

At the end of the annotation process, we were able to compile 27 102 different annotations resulting in a final dataset of 7 395 labeled target-response pairs. Of this, 2 715 (36.71%) were consistent across all questions (*super-consistent* pairs) and the remaining 4 680 (63.29%) were inconsistent in at least one of the main questions and required arbitration. Of this final group, 703 (9.50%) were flagged as "Skipped" and resolved in a group setting by 2 or more *Arbiters*. The proposed methodology allows for different ways of assessing the reliability of the final labels. In Table 5.4, we present the inter-annotator agreement metrics for the different questions and at different aggregation levels. We use Cohen's κ coefficient as its a more robust measure than simple percent agreement calculations, given it accounts for chance agreements. As expected, the hardest question corresponds to the first component of the conversation stance detection task (question 4.a). Even with the inclusion of the two extra categories (the "Explicit" and "Implicit" qualifiers), the annotator agreement is on par or slightly better than other crowdsourced annotated datasets [141]. However, when considering the comparable 4-class task (the majority ignores the qualifiers), we see a significant increase in the agreement between annotators.
The label distributions for the main stance questions are presented next. In Table 5.3, we provide the distribution, by response type, of the Target Stance classification questions. As shown, most of the responses show negative stances toward their government (question **3.b**), which is consistent with studies of the region during the protests [125]. Interestingly, in the case of the Original tweets (question **2.b**), the distribution is more uniform. This can be accounted for by the increased likelihood (due to the sampling strategy) of sampling viral contentious tweets which included tweets authored by numerous verified figures like government officials, supportive politicians and local members of the press. The latter, as shown in Chapter 3, were more

likely to be supportive of their local government during the protests.

	Reply			Quote		
	In Favor	Against	Neutral	In Favor	Against	Neutral
2.a	627	1140	1063	622	623	1101
2.b	1128	1105	1022	553	1168	1023
3.a	848	593	1309	757	468	1079
3.b	315	1898	986	258	1707	756

Table 5.3: Results for the Target Stance classification task for the Response and its target. The "Does not apply" category is omitted for economy of space.

The results for the two parts of the Conversation Stance classification task are presented in Table 5.5. As expected, when compared to similar datasets [141], the sample produced annotated examples are biased towards non-neutral responses. It is worth noting that direct replies are more likely to exhibit disagreement than quotes.

5.4.2 Baseline Results

In this section we discuss baseline results for different classifiers, both feature-based and deep learning architectures, trained on the final dataset. The results correspond to a held-out test set of 950 pairs from the annotated dataset, chosen by sampling users at the root of the conversation tree and ensuring that they are not seen during training (at least as target users). We include several feature-based classifiers given their strong performance in small-scale datasets in general, and in stance classification in particular [75]. These include: a 3-layer Multi-layer Perceptron (MLP), Support Vector Machines (SVM), Logistic Regression (LR) Multinomial Naïve Bayes (MNB) classifier and, a Random Forest (RF) classifier. The features extracted were based on TF-IDF counts of the tokenized terms. Since there were several hyperparameters to be tuned and optimized, a grid-search-validation strategy was used during training to find the optimum set of

	1.b	**2.a**	**2.b**	**3.a**	**3.b**	**4.a (6 classes)**	**4.a (4 classes)**	**4.b**	**4.c**
(%) Revised with Majority	95.13	92.61	92.22	90.38	90.51	86.03	86.00	88.35	92.18
Arbiter κ	89.1	91.52	92.98	90.68	88.71	83.91	86.68	81.52	79.48
Annotator κ	79.06	73.02	74.42	70.91	73.12	69.73	76.88	73.99	71.9
Total κ	79.07	78.63	79.37	76.98	77.14	70.89	75.71	74.98	73.27

Table 5.4: Inter-annotator agreement at three aggregation levels measured by Cohen's κ coefficient. All agreement metrics are compared against the majority label. The *Arbiter* level is limited to "problematic" pairs that achieved a majority after arbitration, the *Annotator* level is limited to *super-consistent* pairs, and the total level aggregates both. The (%) Revised with Majority column measures the percentage of "problematic" pairs that had a simple majority for a given question after being revised by an Arbiter.

(a) Part one (question 4.a).

	Reply	**Quote**
Comment	741	877
Explicit Disagreement	1556	753
Implicit Disagreement	693	375
Ask for additional information	170	168
Explicit Agreement	276	478
Implicit Agreement	546	762

(b) Part two (when applicable).

	Reply		**Quote**	
	Yes	No	Yes	No
4.b	1093	1156	565	563
4.c	292	530	548	692

Table 5.5: Results for the two components of the Conversation Stance classification task.

hyperparameters, using an independent validation set of 10% of the training data. We also include 2 BERT baselines additional to the one described in this chapter. Namely, the first version of the model introduced in Chapter 4 and *TwilBERT* [62]. This latter version is the only other currently available Spanish BERT variant trained on Twitter data, but has the disadvantage of being trained in less data, is mostly focused on European Spanish and does not apply the RoBERTa pretraining framework [86].

In Table 5.6, we present the baseline results for the main questions covered for each annotated pair. Given the asymmetry of the label classes for the different questions, we do not only focus on classification accuracy but also in Macro F1-score. The first thing to note is that, consistent with the annotator statistics presented before, the two questions related to the identification of the event present the least amount of challenge across all classifiers. The "Country Prediction" task encodes question **1.a** into discrete categories, and pairs corresponding to multiple countries were aggregated in a separate category. Approximately 5% of pairs fell to this category and much of the drop of the F1-score, with respect to the accuracy, corresponds to poor performance with this label (this was observed across all classifiers). We note that the two target stance classification tasks are handled as a tweet-level classification problem. As such, the labels for the original and response tweets are constructed by aggregating answers for **2a** and **3a** for the Protest stance and **2b** and **3b** for the Government stance. As shown, the best classification results are consistently observed with *twBETO v1* with the RPT pretraining process, which significantly outperforms the simple feature based models used. We can also see that all *twBETO* variants significantly outperform the *twilBERT* model. This is consistent with the degraded performance noted by the authors when dealing with non-European Spanish variants and the design choice to adopt the traditional BERT pretraining approach (instead of the RoBERTa objective).

The results for the two components of the stance in conversations classification problem are also presented in Table 5.7, for the different Transformer-based baselines. We exclude our simple feature-based variants given their poor performance and for economy of space. The first component, shown in "Conversation Stance" part of the table, presents the results for the original 6-class label defined by question **4.a** and a 4-class variant obtained by ignoring the "Explicit" and "Implicit" qualifiers. Given the increased context necessary for the classification of the response, this task is significantly harder than target stance classification, which is consistent with the annotator agreement statistics. We see that the improvements obtained by the second version

	Country Prediction		Protest Relevance		Protests Stance		Government Stance	
	Accuracy	Macro F-1	Accuracy	Macro F-1	Accuracy	Macro F-1	Accuracy	Macro F-1
Linear SVM	81.0	69.0	79.1	60.0	55.1	48.0	59.2	51.0
Logistic Regression	83.0	72.0	79.6	61.0	57.0	54.0	60.7	55.0
MLP Classifier	82.0	69.0	78.0	59.0	49.4	53.0	55.2	55.0
Multinomial NB	80.0	67.0	70.0	61.0	55.1	53.0	58.8	52.0
Random Forest	80.0	65.0	71.0	60.0	56.6	49.0	59.4	48.0
twilBERT [62]	89.3	79.9	83.2	75.4	60.1	55.2	61.9	57.2
twBETO v0	94.8	85.6	91.8	80.7	63.9	62.6	70.9	65.6
twBETO v0 with RPT	95.2	86.9	93.1	84.9	65.7	64.2	71.4	66.8
twBETO v1	95.5	87.2	93.0	85.8	65.9	65.0	71.6	66.2
twBETO v1 with RPT	**95.7**	**87.3**	**93.9**	**86.0**	**66.2**	**65.3**	**72.0**	**67.5**

Table 5.6: Accuracy and Macro F1-score, in percentage points, for the event identification, target stance classification tasks. The latter is done at a tweet level, not a pair, so the labels for questions **2x** and **3x** are aggregated.

of *twBETO* are more pronounced in both these tasks, which require a pair of inputs. This can be explained by the addition of more conversational data (with the aforementioned Agreement-Disagreement collection) and the improvement of the tokenization method. We can also leverage the two qualifiers to achieve a better understanding of the instances where the model tends to make mistakes, noting that the models are more prone to miss-classify "Implicit" agreements or disagreements as they are more context dependent. Using the baseline obtained with the *twBETO v1* model and, in the case of the best performing class ("Disagreements" with 77% of predictions corresponding to the true label), we note that 58% of miss-classifications correspond to comments, followed by 22% corresponding to Implicit types of Support. Similarly, for the worst performing class ("Comments"with only 54% of predictions correctly assigned), we note that 51% of miss-classifications correspond to implicit forms of disagreement or agreement (with 30% corresponding to implicit types of support).
The second component of this problem is encapsulated in a "Perception of Veracity" task, which combines the labels for questions **4b** and **4c**. This can be done without issues as, given the logic implemented in the each questionnaire, both questions are mutually exclusive. However, there are two ways we construct the negative class for this task, the 4-class variant aggregates the "No" responses to questions **4b** and **4c** in a single category, while the 5-class task does no aggregation. As shown in the table, classifiers for this task show the worse performance, which can be accounted by smaller percentage of positive instances included in the sample.

Leveraging "Implicit" and "Explicit" Qualifiers We can also leverage the extra information provided by the qualifiers to improve the quality of the predictions made on the first part of the conversation stance detection task. Similarly to what was done with the annotator agreement statistics, we can transform our 6-class predictions to the 4-class task by aggregating both types of agreements and disagreements (simply ignoring the qualifiers). In Figure 5.4, we present the column-normalized confusion matrices for both predictions. We note that this process increases the final test macro F1-score of the model by 1.3 percentage points at the cost of 0.7

	Conversation Stance				Perception of Veracity			
	4-classes		6-classes		4-classes		5-classes	
	Accuracy	Macro F-1	Accuracy	Macro F-1	Accuracy	Macro F-1	Accuracy	Macro F-1
twilBERT [62]	54.7	50.1	49.8	44.5	46.1	36.4	42.1	40.7
twBETO v0	61.0	57.2	52.5	48.7	52.2	50.6	48.0	44.8
twBETO v0 with RPT	61.6	57.5	53.9	50.9	52.8	50.5	49.8	46.4
twBETO v1	65.8	62.0	53.3	50.0	53.5	51.6	52.0	49.5
twBETO v1 with RPT	**66.5**	**63.2**	**56.1**	**54.4**	**54.9**	**52.7**	**53.7**	**50.5**

Table 5.7: Accuracy and Macro F1-score, in percentage points, of the *twBETO* baseline for the conversation stance classification tasks. The "Perception of Veracity" task combines the labels, that are by construction mutually exclusive, of questions **4b** and **4c**.

percentage points in accuracy. We also note that the model is effective at distinguishing between disagreements and agreements, as only 6% of the former type of predictions corresponds to true agreements, while in the case of the latter, this error rate increases to 10%.

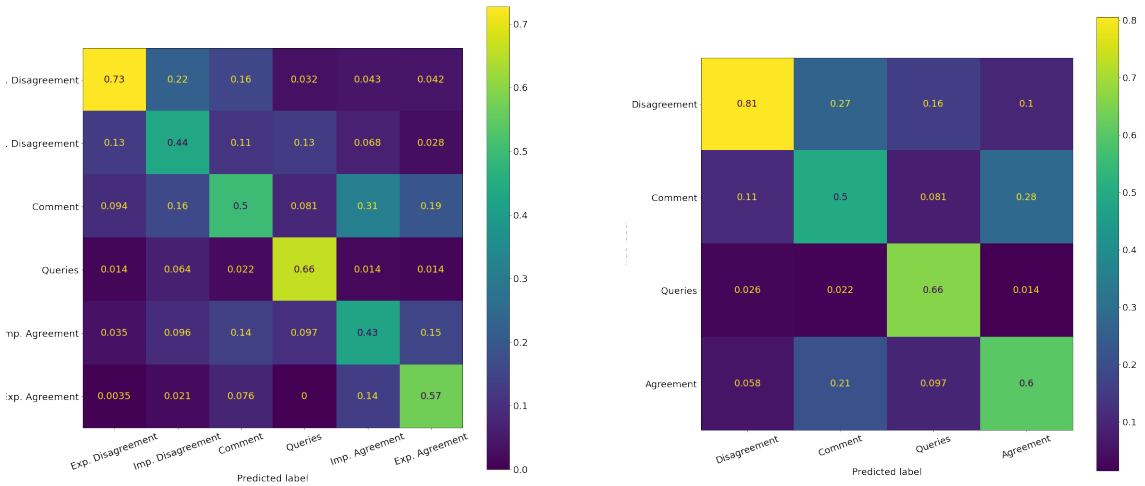

Figure 5.4: Column-normalized confusion matrices, on the test set, for the 6-class conversation stance predictions (left) and their corresponding 4-class results after ignoring the qualifiers (right). This methodology achieves a 66.3% accuracy and 64.5% F1-score on the test set, improving the F1-score obtained by training directly on the 4-class labels.

5.4.3 Consolidating the Conversation Stance Detection Task

The main purpose of the classifiers built in this work is to identify responses that (dis)agree with their target and whether they affirm their target is spreading falsehoods or is telling the truth. In

this section we explore a simple method to increase the precision of both models' predictions in these key classes by consolidating their results. That is, for example, we want to maximize the chance that when the "Perception of Veracity" model predicts a response to be labeling its target of spreading falsehoods, it is indeed the case (even at the expense of missing other true responses of this type). We will work with the consolidated 4-class "Conversation Stance" and the 5-class "Perception of Veracity" classifiers, and apply a simple heuristic to make the models more conservative in their predictions of these classes. This is based on a model averaging approach and the steps taken are as follows:

1. For the consolidated 4-class "Conversation Stance" (4-class C.S.) classifier:
    - When the 5-class "Perception of Veracity" (5-class P.V.) predicts a response to be "Neutral", assign a "Comment" label.
    - When the model predicts "Agreement" and the 5-class P.V. does not predict agreement ("Normal Agreement" or "Labels Truth"), assign a "Comment" label.
    - When the model predicts "Disagreement" and the 5-class P.V. does not predict disagreement ("Normal Disagreement" or "Labels Falsehood"), assign a "Comment" label.
2. For the 5-class "Perception of Veracity" (5-class P.V.) classifier:
    - When the model predicts "Labels Truth" and the 4-class C.S. classifier does not predict "Agreement", assign "Neutral".
    - When the model predicts "Labels Falsehood" and the 4-class C.S. classifier does not predict "Disagreement", assign "Neutral".
    - When the 4-class C.S. classifier predicts "Comment" or "Queries for more information", assign "Neutral".

In Figure 5.5 we present the results of applying this heuristic on both classifiers. In the case of the "Conversation Stance" classifier, even though there is a 1 percentage point reduction in overall accuracy we are able to improve the "Disagreement" precision by 2 points and the "Agreement" precision by 9 points. For the "Perception of Veracity" classifier we see more conservative improvements, with only 2 percentage points increases for both the "Labels Truth" and "Labels Falsehood" classes. It is important to note that both models are effective at distinguishing between agreements and disagreements or responses that label their targets as spreading falsehoods or telling the truth. The worst case occurs with the "Agreement" class, were 7% of the time the model incorrectly assigns this label to a true disagreement. We will rely on this more precise version of both models for the analysis undertaken in the next chapter.

5.5 Limitations

The scope of the dataset constructed is limited to tweet pairs that occur in the first level of a conversation tree. As noted by other studies on the classification of stance in conversations [144], this can degrade the performance of models trained on this data and used to predict instances that occur in lower levels of a conversation tree, as the context of the interaction is less dependent on what is observed in the pair. Moreover, the conservative classifiers for stance classification

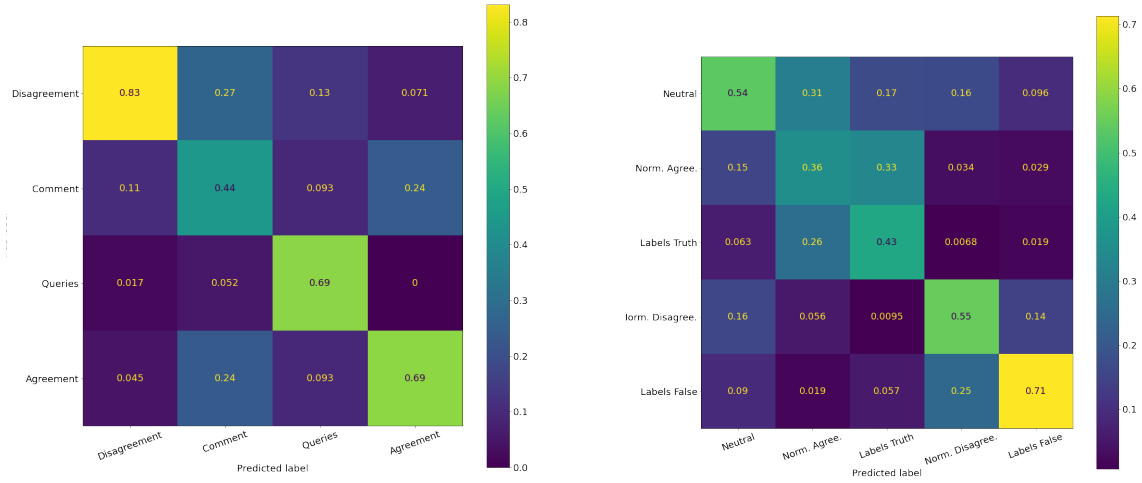

Figure 5.5: Column-normalized confusion matrices, on the test set, after the model averaging heuristic for the 4-class "Conversation Stance" classifier (left) and 5-class "Perception of Veracity" classifier (right). In the case of the former, we achieve a 2 percentage point improvement in the "Disagreement" precision and a 9 point increase in the precision of the "Agreement" class. For the latter, we achieve a 2 percentage points increase in the precision for both the "Labels Truth" and "Labels Falsehood" classes.

introduced in this chapter opted for a simple model averaging approach, instead of more explicit forms of multi-task learning. Although this was deemed sufficient given our focus on improving the precision for key classes, even at the cost of overall performance, a joint optimization of the losses for each task (in a multi-task setting) can be an effective way to learn more robust features. Future work should leverage the multi-task nature of the constructed dataset to improve the overall performance of the models trained.

5.6 Discussion

In this work, we hand-labeled the target and conversation stance of a dataset of 7395 target-response pairs. We focused on Spanish Twitter conversations around the protests that shocked the South American Region at the end of 2019. The sampling methodology focused on contentious conversations and stratified interactions based on whether they are Replies or Quotes. To the best of our knowledge, this is the first annotated dataset of its type in Spanish and one of the largest available for this task. Importantly, we propose a unified framework for the annotation of data for the detection of stance in conversations. First we identify agreements, disagreements, or neutral responses; and then we identify whether non-neutral responses are supporting (denying) the veracity of their target. We believe that this unified treatment can help bridge these two related research areas and is necessary in order to explore how polarization in online discussions

can affect the spread of rumors.

Chapter 6

Towards a policy-oriented test-bed for the spread of Contentious Messages in Twitter

6.1 Introduction

In recent years, there has been a renewed interest in the spread of disinformation on online social media, its effects on public discourse, and how it relates to the increasing levels of polarization reported on different platforms. In academia, much of the effort has been spent characterizing the salient traits of the diffusion process of false information, especially as it relates to other types of information. For example, Vosoughi et al. [126] contrasted the spread of true and confirmed false news stories on Twitter, finding that false articles (particularly as they relate to political events) spread farther than their counterparts. However, there is little understanding on the effect that this has on society or its institutions and what capacity do malicious actors have to orchestrate elaborated campaigns that inflame these negative effects. More importantly, much work remains to be done to explore viable avenues for policy that would curtail the spread of disinformation or mitigate its pervasive impacts. Lazer et al. [79], posit two venues of intervention to increase community resiliency, those that seek to educate and empower individuals in their interactions with social media and those that promote structural changes in the platforms to limit user's exposure to false information. Some work has been done to explore the former, by investigating community-based options for fact checking or shaming and reporting bad actors [7, 119]. However, to the best of our knowledge, no work has yet to explore effective ways to design policies that implement these interventions and would allow the evaluation of their effectiveness under different scenarios. To design and implement effective and efficient interventions requires a more detailed understanding of the determinants of community resiliency to the spread of disinformation, specially, during polarizing events. In particular, as we showed on Chapter 2, how the members of polarized communities interact with each other and with outsiders holding opposing views, can affect how false information and the response to it spreads through these communities.

In order to begin to develop a viable testbed for the aforementioned policies, we propose an agent-based dynamic-network model of contentious Twitter conversations. The proposed model augments an existing and validated Twitter Simulation based on CONSTRUCT's model for in-

formation diffusion [35]. This type of simulation is determined by agents that form and sever network ties over time as a function of their properties and not their position on a grid [26]. We validate the proposed simulation via a case study focused on the "cacerolazo" (pot and pan banging) protests that took place during government-mandated curfews in the height of the 2019 Ecuadorian protests. This event is specially relevant to the themes explored in this work, as people both supporting and opposing the government partook in it, while confined to their homes, and actively claimed in Twitter that it supported their position. Moreover, the smaller scale of this event, when compared to the universe of data collected for the protests, allows the simulation of a non-trivial part of the discussion. Through the usage of the classifiers developed in the previous chapters, we are able to show that the observed levels of polarization in the way that users shared information is not consistent with the "filter bubble" explanation of polarization. Moreover, we show that users were actively exposed and interacted (mostly negatively) with content produced by the other side of the argument. This implies that the observed phenomena is more consistent with social media practices that reflect polarized content engagement. This is important as policy interventions that focus on changing the way that content is shown to give access to opposing views, without any other type of input or moderation, are not likely to be an effective method for reducing polarization.

6.2 Background

6.2.1 The Ecuadorian "Cacerolazo"

There is a long and storied history of the usage of pots and pans to protest local governments, dating back to France in the 1800's [28]. Modern iterations of this type of protest have largely been confined to the Latin American region and Spain, where they are respectively known as *cacerolazos* and *caceroladas*. The first instances of this type of dissent in the region where reported in Chile during the 1970s, while they became an important form of protesting in Argentina during the early 2000s. The 2019 South American protests were no exception, as numerous *cacerolazos* were observed in Ecuador, Chile and Colombia.

The first of these events took place in Ecuador on the night of October 12, after 10 days of generalized protests and the temporary movement of the seat of Government from Quito to Guayaquil (see Section 1.2.2 for further details on the protest). The event was convoked mainly through social media by citizens not directly related to the parts in dispute, largely because of the mandatory curfew that was in place. The decentralized origins of this *cacerolazo*, also gave way to a double interpretation of the intentions behind the manifestation. While the event trended on Twitter, discussions between users actively disputed whether the cacophony of sounds that filled the streets were a rebuke of the government or a call for peace and the end of the protests. In Figure 6.1, we present an illustrative interaction between a major Ecuadorian news outlet ('El Universo"), which unequivocally claimed that the event was a call for peace, and two other users. The first user disputes this claim, accusing media organization of spreading disinformation, while the other sides with the event as reported. It is worth noting that, as explored in Chapter 3, the country's mainstream media was largely supportive of Moreno's government during the protest as it had struck down several laws put forth by the previous administration which constrained the

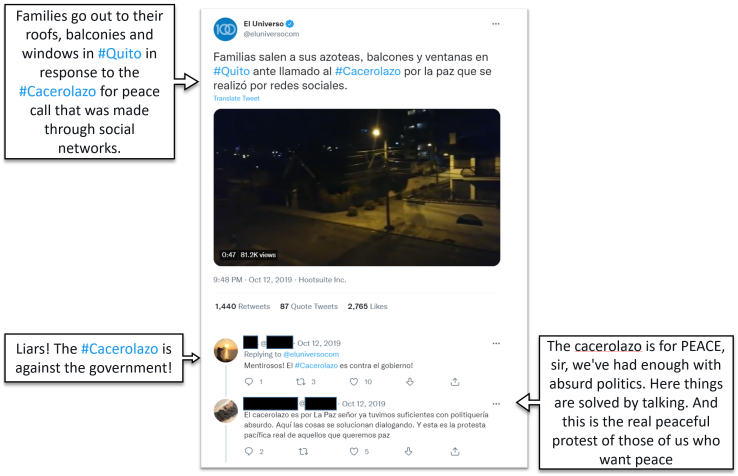

Figure 6.1: Contentious Twitter conversation discussing the motivations of the Cacerolazo that took place in Quito, during the 2019 Ecuadorian protests. The Origin tweet is a report from "El Universo", a major Ecuadorian news media outlet.

operations of the press.
In this work, we focus on the Twitter conversation of the *cacerolazo* of October 12, collected by matching any conversation that presented a tweet containing the substring "cacerola" in a 1-day window of this date. This resulted in 60 thousand original tweets (not including retweets) from 6.1 thousand users that either replied/quoted or were replied/quoted by another user who posted for this event. We then extracted and additional 222 thousand retweets between these users in the time span analyzed, ensuring that both retweeter and retweeted parties were in the identified set. After applying the classifier developed in Chapter 4, we determined that 4 127 (67.0%) of users involved, in this connected core of the overall conversation, were against the government.

6.2.2 Simulation of Twitter Discussions

Although numerous works have modeled information diffusion in an online medium (mostly Twitter) [58, 95, 107, 110], few consider the effect that the platform has on the formation of user networks and their interaction . These environments, however, impose a series of rules for interaction and content promotion that have an important effect on observed network topologies and the diffusion of information [13, 81]. For example, Weng et al. [128] showed how limits on the size of a user's timeline is sufficient to recreate the heterogeneous diffusion of memes in Twitter. This introduces an economy of attention, which imposes competition among different topics discussed at a given moment. The Twitter environment introduces four main forms of interaction between users: retweets, direct replies, quotes and likes. The later is generally not considered in scientific studies of the platform, even though is the most prominent, as is not publicly available through the API. The retweet action implies a forwarding of the observed tweet to the user's followers which, as we show in Section 6.4.1, generally implies support of the content. On the other hand, direct replies follow different broadcasting rules as they only appear in the timeline of other users if they follow both the target and the replier. Quotes are hybrid between replies and

retweets that were introduced by Twitter in 2015 in response to user's tendency to 'manually' retweet. Quotes are broadcasted as if they are original tweets, which implies that they are only shown to the followers of the user and does not appear in the timeline of the quoted user (which is not the case for direct replies). The usage of this new feature has gradually been increasing, and has helped increase political discourse and its diffusion [45]. Finally, users also have the capacity to modify their social network by following/unfollowing other users. A recent framework for Twitter simulations incorporated these agent decisions based on the mechanisms established in the CONSTRUCT model for learning and information diffusion [26]. Here, agents update their social connections based on homophily with other user's follower network and their shared belief of different knowledge bits. In this work, we build upon this framework by augmenting the cognitive model of users to also include their stance for any knowledge they posses. In this way, the incentives of a user to respond positively (negatively), increase as her beliefs correlate more positively (negatively) with the relevant message (we introduce confirmation bias). In this we deviate from the standard Twitter Interaction model used in CONSTRUCT and in other frameworks [128], where all messages have a uniform probability of being replied to. As we show in this work, this mechanism is sufficient to recreate the different forms of polarization observed during the event, which is not possible with the current implementation of Construct's model for Twitter.

6.3 Methods

6.3.1 Quantifying Observed Polarization

In this work we explore the different facets of the polarization observed during the *Cacerolazo* protests. For a more formal evaluation of this phenomena, we consider two metrics commonly used to measure polarization and controversy in social networks:

- External-Internal Ratio ($E|I$ Ratio) [74]: a social network measure of the relative density of internal connections within a social group compared to the number of connections that group has to the external world. For a given group, we can define this ratio as:

$$E|I \text{ Ratio} = \frac{E - I}{E + I}$$

  Were E and I are the number of edges pointing to nodes outside and inside the group, respectively. It ranges from -1 (all edges are internal) to 1 (all edges are external) and is not robust to asymmetries in the sizes of the groups studied.

- Random Walk Controversy (RWC) metric [46]: it is only defined for two groups and intuitively captures the notion of how unlikely a user from either side is to reach an authoritative user from the other group. It is defined as:

$$RWC = P_{XX}P_{YY} - P_{XY}P_{YX}$$

  Where P_{AB} $A, B \in \{X, Y\}$ is the conditional probability that the walk starts in A and ends in B. Walks end when a user reaches a predefined high-degree node (computed

independently for each side). This metric is generally estimated via Monte Carlo methods and has the advantage that is not skewed by asymmetries in the size of the groups nor in the degree of the nodes.

6.3.2 Simulation Description

The simulation is developed using Construct API[1], which is an agent-based simulation framework based in C++. We expand the current implementation of the CONSTRUCT agent-based modelling platform [35] to augment both the standard *Twitter Interaction Model* and its abstraction of social media user's behavior. The purpose of this exercise is to extend the represented agent's cognitive capabilities to include a stance towards the singular knowledge bits embedded in a media event (CONSTRUCT's standard abstraction of the information contained in a particular tweet). This module also introduces a mechanism simulating a user's observed tendency for social media practices exhibiting polarized content engagement. As we discuss in Section 6.4.1, the different Twitter interaction networks we study during the event manifest the type of polarization consistent with confirmation bias on part of the users.

For our current purposes, we instantiate a fully-fledged Twitter model incorporating all the standard networks that come within CONSTRUCT's implementation and introduce new ones to meet our goals. We define two new classes within the environment, namely **Stance_Social_Media** and **stanced_media_user** classes that extend the standard **Social_Media** and **media_user** used by the model. These new classes maintain the original modelling schedule of the Social Media Interaction model but extend it's functionality by adding a new *knowledge stance model*. As part of this extension, two new networks are included, the *knowledge stance network* and the *knowledge stance transactive memory network*. Based on these additions, the main changes occur on the underlying logic on how stance-aware agents decide when to create a tweet, or how they interact with content produced by other users. This includes changes in the way an agent decides to follow other users, or which tweets they decide to reply, quote, or retweet.

Our *Stance Twitter Interaction Model* includes both the *knowledge network* and the *twitter follower network* at initialization. As it happens with the standard Twitter model, users spontaneously create original tweets which may or may not mention other users (represented as media events containing a single piece of knowledge). The number of tweets a user will generate in a time period is equal to an agent's *post density* attribute times the *interval time duration* parameter. When a tweet is generated, a random piece of knowledge is selected and the tweet's author attaches their trust and stance score for the piece of knowledge[2]. The stance score is a value within the range $[0, 1]$ that defines each agent's position relative to one of three potential ideological clusters (agreement, disagreement, and neutrality) towards each knowledge bit. For our purposes, we say that she is in disagreement if $S \in [0, 0.4]$; neutral if $S \in (0.4, 0.6]$; and, in agreement if $S \in (0.6, 1]$. Agent i's trust of a piece of knowledge k is stored in the *knowledge trust network* (denoted $T_{i,k}$), and her stance is stored in the *knowledge stance network* (denoted $S_{i,k}$).

[1] https://github.com/CASOS-IDeaS-CMU/Construct-API

[2] As in the standard model, trust represents how likely an agent/user believes a piece of knowledge is factually true and is in the range $[0, 1]$.

When agents read an event, their base probabilities for creating replies, retweets, and quotes are given by their *reply probability*, *repost probability* and *quote probability* respectively, and a response always refers to the same knowledge bit as the original tweet. Retweets attach the same trust and stance score as the parent event (a simple forwarding mechanism). Replies and quotes contain the responding user's trust and stance score of the parent event's piece of knowledge.
However, our implementation differs from the standard CONSTRUCT model in how a user's stance towards the read knowledge bit, influences the probability of each type of action. That is, we include a behavioral mechanism reflecting user's tendency for confirmation bias. For instance, the probability of a retweet follows the subsequent behavioral pattern depending on the corresponding entries in the $S_{i,k}$ and $S_{j,k}$ matrix for both the ego i and alter j agents who generated the tweet:

- If both $S_{i,k}, S_{j,k} \in (0.4, 0.6]$ (that is, both agents have a neutral stance towards knowledge bit k), then the retweet probability just equates the corresponding base probability, denoted RP_i^{base}.
- If both $S_{i,k}, S_{j,k} \in [0, 0.4]$ or $S_{i,k}, S_{j,k} \in (0.6, 1]$ (that is, both agents share the same type of stance towards knowledge bit k) then the retweet probability is re-scaled via the following formula:

$$\text{RP}_i = \left(RP_i^{max} - RP_i^{base}\right) \frac{0.4 - |S_{i,k} - S_{j,k}|}{0.4} + RP_i^{base}$$

  where RP_i denotes the adjusted retweet probability for agent i and RP_i^{max} denotes the maximum repost probability, which is given on agreement, which is given by the *max repost probability on agreement* parameter. Note that, given the symmetry between the assumed size of the agreement and disagreement spaces, the 0.4 value serves to scale their stance difference to the size of the corresponding quadrant.
- For any other case, both agents disagree on their stance towards knowledge bit k, and the appropriate formula depends on user i stance cluster. That is:

$$\text{RP}_i = RP_i^{base} - \left(RP_i^{base} - RP_i^{min}\right) \frac{0.4 - |S_{i,k} - l|}{0.4}$$

  where RP_i^{min} denotes the minimum retweet probability, which occurs on disagreements, and is given by the *Stance Twitter min repost probability on disagreement* parameter. The l constant denotes the appropriate stance cluster limit and is equal to 0.4 if $S_{i,k} \in (0.6, 1]$ to 0.6 if $S_{i,k} \in [0, 0.4]$.

For the reply and quote media event types, the corresponding adjusted response probability (RP_i) follows a similar behavioral model with a few adjustments. That is:

- If both $S_{i,k}, S_{j,k} \in (0.4, 0.6]$ (both agents have a neutral stance towards knowledge bit k), then the repost probability just equates the corresponding base probability, denoted RP_i^{base}.
- If both $S_{i,k}, S_{j,k} \in [0, 0.4]$ or $S_{i,k}, S_{j,k} \in (0.6, 1]$ (both agents share the same type of stance towards knowledge bit k) then the probability is re-scaled via the following formula:

$$\text{RP}_i = \left(RP_i^{max_{agree}} - RP_i^{base}\right) \frac{0.4 - |S_{i,k} - S_{j,k}|}{0.4} + RP_i^{base}$$

where RP_i denotes the corresponding response probability for agent i and $RP_i^{max_{agree}}$ denotes the maximum probability on agreement, which is given by the *max reply probability on agreement* or the *max quote probability on agreement* parameters depending on the type of response.

- For any other case, both agents disagree on their stance towards knowledge bit k, and the appropriate formula depends on user i stance cluster. That is, if $S_{i,k} \in (0.6, 1]$ then:

$$\text{RP}_i = \left(RP_i^{max_{disagree}} - RP_i^{base}\right) \frac{0.4 - |S_{i,k} - l|}{0.4} + RP_i^{base}$$

where $RP_i^{max_{disagree}}$ denotes the maximum response probability on disagreement, which is given by the *max reply probability on disagreement on disagreement* or the *max quote probability on disagreement* parameters depending on the type of response. As before, the l constant serves to scale the probability to the the appropriate stance cluster limit.

Finally, after reading an event, the user can decide to follow/unfollow the event's author. As in the standard model, a defining parameter for this decision is the event author's attribute *charisma*, denoted C_j. The attribute is ranged in the $[0, 1]$ interval and represents the base probability that the reading user will follow the event's author. As previously indicated, a final contribution we incorporate to CONSTRUCT is the *knowledge stance model*, which updates the user's a *stance score* $S_{i,k}$ to knowledge items based on an agent's position towards a knowledge bit represented by the *knowledge stance network*. When agents learn a new knowledge, the corresponding stance is set to 0.5 (neutral) and the addition of stance does not affect the parsing of the knowledge item. Knowledge stance is also parsed by this model and that information is added the *knowledge stance transactive memory network*. An agent's stance is then updated during the **Clean Up** function based on the stance of the agent's alters in their stance transactive memory. Agents gradually change their stance on a piece of knowledge based on the *stance relax rate* parameter (denoted r_s) which is a value in the range $[0, 1]$. The updated knowledge stance $S'_{i,k}$ is thus computed by:

$$S'_{i,k} = (1 - r_s)\, S_{i,k} + \frac{r_s}{\left|\text{KSTM}^*_{i,j}\right|} \sum_{k \in \text{KSTM}^*_{i,j}} \text{KSTM}_{i,j,k}$$

Where $\text{KTSM}_{i,j,k}$ denotes the entry from the *knowledge stance transactive memory network*, $\text{KTSM}^*_{i,j}$ is the set of items agent i has in their knowledge trust transactive memory for agent j. A similar mechanism is currently in place in CONSTRUCT's Twitter Interaction model to update the trust of users to the different bits they currently know.

Initialization of the simulation

The backbone of the Twitter Interaction model, and the main ingredient required for non-trivial simulations, is the Follower/Friend network. In this network, two users are connected by a directed edge if one follows the other. Given that our focus is to replicate the behavior observed during the event, we collected this network for the set of identified users[3]. However, given the

[3]The Friends network was crawled more than a year after the studied events transpired and, as such, it might not fully reflect the true network at the time.

computational cost required to simulate the whole conversation, we opted for a sample of 10% of the users obtained via a directed random walk with induced graph sampling. This sampling method has been found to be effective at reproducing network topological features [82], and in our case retains a similar level of polarization (with a difference in the RWC score of only 0.05). The original and sampled networks are presented in Figure 6.2, and, as we discuss in the following section, they do not show significant levels of polarization.

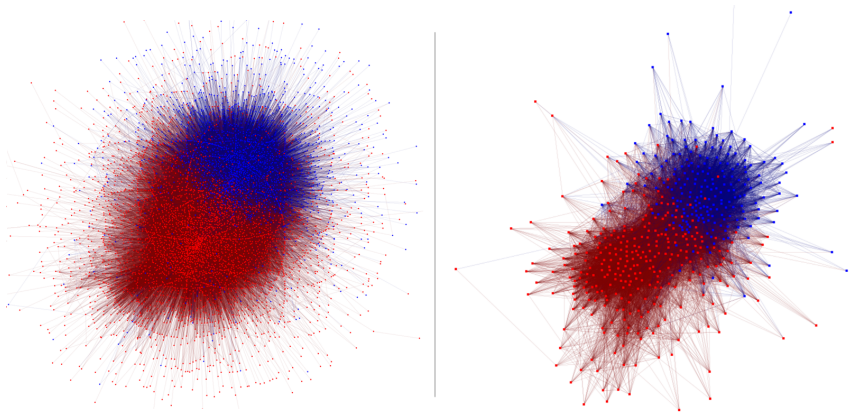

Figure 6.2: Friend/Follower network for the users who participated in the *cacerolazo* protests. The left panel shows the complete network and the right panel to the 10% sample obtained via random walk induced graph sampling. Nodes are colored based on their predicted government stance: red corresponds to "Against" and blue to "In Favor".

The final ingredient required for initialization is to define the knowledge space. In this work we opt for a simple abstraction to reflect the partisan nature of the event. That is, we assume that there are 60 different knowledge bits divided in three groups of equal size, where they are against the government if $k \in [0, 20)$, neutral if $k \in [20, 40)$ and in favor of the government if $k \in [40, 60)$. We assume an average density 0.30 for the Knowledge network, which implies that users "know" 30% of the knowledge space. This information is then used when initializing the Knowledge Stance and Trust networks, as the corresponding value is assigned randomly based on the stance of a given user. We assume that "Pro" users "agree" with knowledge bits that match their stance, are "neutral" to neutral knowledge and "disagree" when the knowledge opposes their stance. Hence the stance value for each known knowledge bit is sampled from a Normal distribution with a mean centered at the middle of the corresponding stance space ($0.2, 0.5$ or 0.8) and a standard deviation of 0.05. Trust towards known bits follows a similar logic, with the exception that we treat neutral knowledge as "facts" and assume that users from both sides of the discussion are likely to "trust" this type of knowledge at initialization. The stance and the other parameters required for the initialization of a given user are either estimated (when possible) or calibrated to match data. These parameters need to be defined at a given time resolution that is consistent with the time scale assumed for each simulation step. In this work, we assume that each time step represents 4 hours and hence the simulation lasts for 12 steps. For more details see Section 6.4.2.

6.4 Results

In this section we describe the main empirical regularities observed during the event and also present the results of the virtual experiments undertaken to recreate them.

6.4.1 Exploring the Conversation Networks

In Figure 6.1, we present two different conversation graphs between the identified users, the Retweet and Response networks. In the former, two users are connected with a directed edge if one user retweeted the other while, in latter, two users are connected by a directed edge if one either quoted (retweet with an added comment) or replied the other. As before, users nodes are colored based on their stance, with red representing opposition and blue support. The first thing to note is that, contrary to the traditional "filter bubble" view of polarized discussions, the response network shows little segmentation based on the stance of the users. This implies that, during the event, users from both sides of the ideological spectrum actively engaged with each other. In contrast, retweets are highly clustered based on the stance of the users, and show the type of polarization that is characteristic of these types of networks during controversial events [46]. We note that most bridging nodes on the blue cluster correspond to verified users and mainstream media organizations (denoted with special symbols) and are retweeted by a minority of users in opposition to the government. In contrast, news organizations identified in the red cluster correspond to small leftist media operations (e.g.: "Prensa Bananera", "kolectiVOZ", etc.), and Russian or Venezuelan regional outlets (e.g.: "ActualidadRT", "TeleSUR", etc.). As we can see, this latter group is firmly within the red cluster as is not retweeted by any supporter of the government. Both of these findings are consistent with the results presented in Chapter 3.

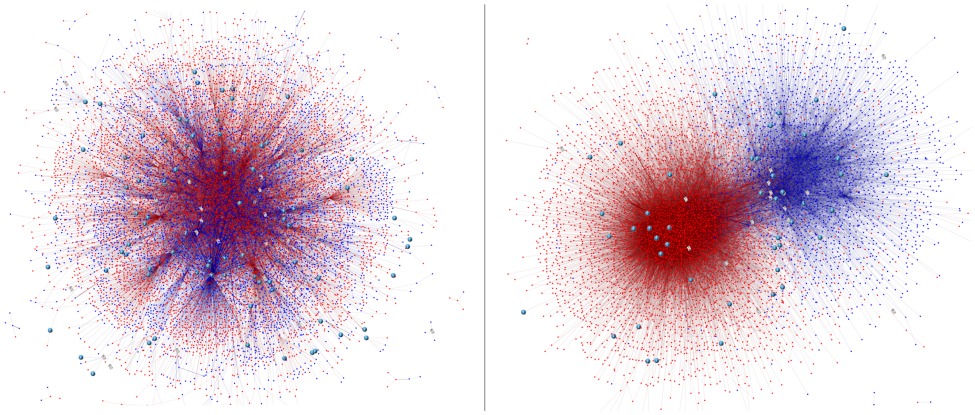

Figure 6.3: Conversation networks for the *cacerolazo* that occurred on Quito on the night of the 12 of October, 2019. The left panel corresponds to the response network (aggregation of tweets and replies) and the right panel to the retweet network. Nodes are colored based on their predicted government stance: red corresponds to "Against" and blue to "In Favor". Nodes corresponding to News Media have a white newspaper symbol while verified users are picture as a circled "i".

In Table 6.1, we present the results of the polarization metrics for the most relevant Twitter

networks collected for the identified users. Consistent with what was noted for Figure 6.3, the Retweet network is highly polarized across all metrics, as most of its edges are internal and the RWC score is above 0.6. On the other hand, the Reply network shows no polarization in any metric, while the Quote network exhibits a mixture of both types of behaviors (which is consistent with its inception in Twitter) but is not significantly polarized. Also, the Friend/Follower network does not show a significant level of polarization, see also Figure 6.2, despite having the majority of its edges be internal for both groups. Albeit, this last measure is skewed due to the asymmetry of the group sizes. The low RWC score can also be explained by users' tendency to directly follow, or to be a short path away, of an authoritative figure in the other group. This evidence suggest that the polarization levels observed during the event are not a result of filter bubbles, as users are actively exposed to content from the other side and their tendency to mostly share content from users of similar views (via retweets) is more consistent with confirmation bias. This is consistent with what has been observed by other researchers, in other social contexts and in English discussions, for Facebook [47, 112] and YouTube [17]. We will next take a deeper dive into the Response network, by estimating the conversation stance of the different interactions.

	Friends	**Retweet**	**Reply**	**Quote**	**Response**
$E\|I$ Ratio (Against)	-0.58	-0.91	0.06	-0.73	-0.24
$E\|I$ Ratio (Pro)	-0.54	-0.70	0.28	0.06	0.22
R.W.C	0.32	0.71	-0.11	0.37	0.02

Table 6.1: External|Internal Ratio and Random Walk Controversy (R.W.C) for different user interaction networks during the event.

Estimating the Conversation Stance

To provide a better understanding of the nature of the responses between users, we are going to leverage the conservative version of the "Conversation Stance" and "Perception of Veracity" classifiers introduced in Chapter 5. This instance of the models is chosen as it maximizes the precision of the estimation for non-neutral responses, even as it comes at the expense of a small drop in overall accuracy. In Figure 6.4, we present two views of the Response network based on the predicted "Disagreements" (18% of the interactions), and "Agreements" (30% of the interactions). We note that the agreement network shows a polarized structure more reminiscent of the Retweet network, which is consistent to the conventional wisdom that retweets mostly signify endorsement. On the contrary, the disagreement network shows considerable mixing of users of different stances, with several in a "star" graph structure with users of different stances actively "disagreeing" with them. As before, in Table 6.2 we present the polarization metrics for the different types of interaction networks based on the predicted conversation and rumor stance of the edges. First note that, for all networks, responses that either disagree or label their targets as falsehood show no polarization, with a majority of edges directed towards users of opposite stances. On the other hand, responses that agree or state the veracity of their target show different behaviors for Quote and Replies, as the polarization observed for the former closely matches the

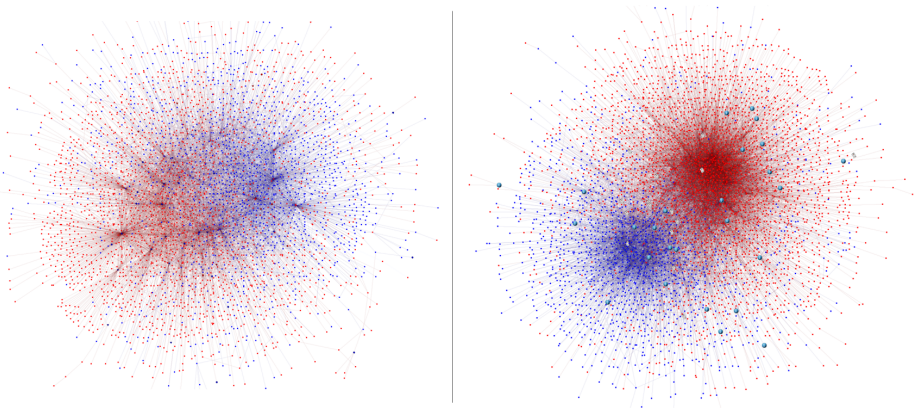

Figure 6.4: Two views of the Response Network (aggregation of tweets and replies) based on the predicted "Conversation Stance" of the interactions. The left panel corresponds to predicted "Disagreements" (18% of responses) and the right panel to predicted "Agreements" (30% of responses). Nodes are colored based on their predicted government stance: red corresponds to "Against" and blue to "In Favor".

	Reply			**Quote**			**Response**		
	$E\|I$ Ratio		R.W.C	$E\|I$ Ratio		R.W.C	$E\|I$ Ratio		R.W.C
	Against	Pro		Against	Pro		Against	Pro	
Agreement	-0.57	-0.47	0.37	-0.88	-0.64	0.62	-0.76	-0.54	0.48
Disagreement	0.49	0.50	-0.26	0.16	0.48	-0.23	0.45	0.50	-0.28
Labels Falsehood	0.55	0.60	-0.33	0.19	0.44	-0.22	0.50	0.58	-0.32
Labels Truth	-0.46	-0.42	0.29	-0.87	-0.61	0.67	-0.76	-0.53	0.50

Table 6.2: Polarization metrics for the different types of user response networks based on the predicted conversation and rumor stance of the edges.

Retweet network. However, in the case of Replies, we observe polarization levels closer to what was observed for the Friend network.

6.4.2 Other empirical regularities

For the initialization of each agent in the simulation, we require to determine a series of parameters capturing their level of activity at the time-scale represented by each step of the simulation. For this reason we explored the distribution of user activity (measured in number of tweets posted) at different timescales, and found that it is well approximated by a power-law distribution. Moreover, we found that the shape of the power-law is not an artifact of the time scale chosen, and is just re-scaled based on the frequency. This is consistent with what have been observed for other empirical studies of Twitter [128]. As an example of this invariance, in Figure 6.5, we show the relationship between the number of Followers/Friends of a user and their average number of posts at a given timescale. We see that the most active users follow between

100 and a 1000 other users, or have around 1000 followers. As before, this relation is not an artifact of the timescale chosen. At instantiation, we assign the *post density* parameter, based on the average number of tweets observed for a given user included in the sample at a 4-hour window (the time-scale chosen for the simulation).

The final element required is the *reading density* and the probability of the different types of

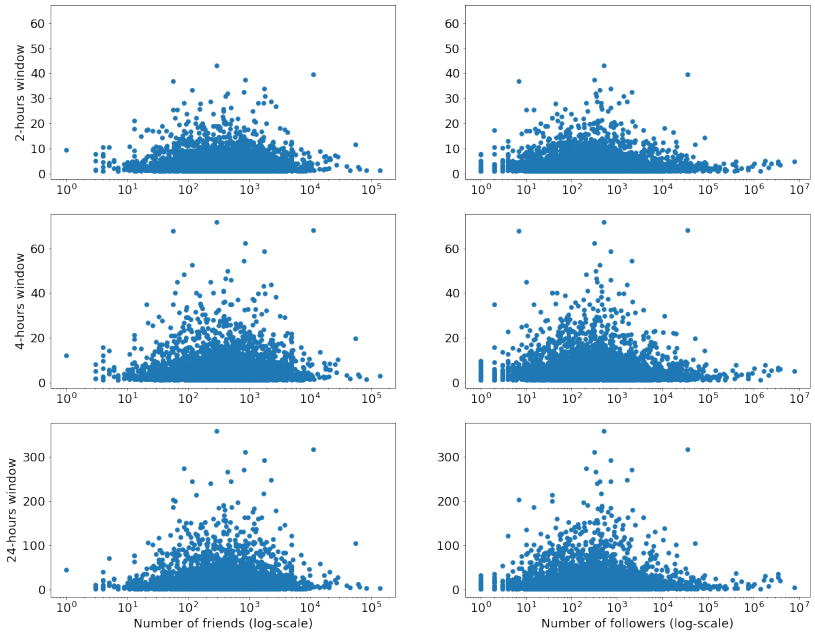

Figure 6.5: Average number of tweets poster by a user as a function of the number of Friends (left) and Followers (right). The number of Friends/Followers is presented in presented in log-scale.

interactions undertaken by each user. However, its is not possible to directly estimate these parameters as the different metrics required to do so are not publicly available. Nonetheless, we are able to establish the relationship between the different types of interactions. For example, the amount of retweets and other responses made by a given user respond to the amount and type of tweets the user read and do not change as a function of the response. In Table 6.3, we present the estimated ratios of the different types of interactions as a function of the maximum retweet probability. As before, these ratios are highly robust to the time scale chosen to calculate them. This allows us to reduce the number of parameters determined via calibration to two, the *reading density* and the *maximum retweet probability*.

6.4.3 Virtual Experiments

In this section we present the results of the different simulations performed both for calibration and to evaluate the proposed stance mechanisms. The main purpose of the proposed experiment is to explore whether the observed levels of polarization observed for the event can be replicated only as a function of the Follower network (via the current implementation of the CONSTRUCT Twitter simulation) or if the proposed Stance module is required. That is, we want to show if the proposed mechanisms are *sufficient* to replicate the behavior observed empirically. The calibration experiments are necessary as some parameters can not be directly estimated based on the data used for this study. The simulation encompasses the days of October 12 and 13 of 2019 (where most of the Twitter activity was seen) and is done for 12 time steps, each representing a 4-hour window. In Table 6.3 we present an overview of the proposed experimental design. Each simulation included 10% of the empirical follower network (600 agents) and lasted in average 3 hours per run. Given the memory requirements of each run, only 3 instances could be done concurrently. In total, the virtual experiment required 144 hours (6 days) of computing time.

Independent Variables	**Values**	**# Test cases**
Reading density	10, 15, 30, 60	4
Max retweet prob. (p^R_{max})	0.09, 0.18, 0.3	3
Inclusion of Stance Module	True, False	2
Control Variables	**Values**	**Source**
r_s and r_t	0.20	CONSTRUCT Default [35]
Charisma	0.50	CONSTRUCT Default [35]
Post density	Empirical Distribution	
Base retweet prob.	0.50 p^R_{max}	Estimated
Base quote prob.	0.01 p^R_{max}	Estimated
Max agreement quote prob.	0.11 p^R_{max}	Estimated
Max disagreement quote prob.	0.04 p^R_{max}	Estimated
Base reply prob.	0.02 p^R_{max}	Estimated
Max agreement reply prob.	0.22 p^R_{max}	Estimated
Max agreement reply prob.	0.10 p^R_{max}	Estimated
Dependent Variables	**Value Type**	**Values seen in data**
RWC of Retweet network	Decimal (0 to 1)	0.71
RWC of Reply network	Decimal (0 to 1)	-0.11
RWC of Quote network	Decimal (0 to 1)	0.37
RWC of Response network	Decimal (0 to 1)	0.02

Table 6.3: Virtual Experiment table for a 4-3-2 experimental design. The simulation was repeated 6 times for each test case, with a runtime of 12 cycles (representing a 4-hour period each) and included 600 agents (10% sample of the follower network).

	Twitter Interaction Model	Stance Twitter Interaction Model
Retweet	0.24 (0.06)	0.86 (0.06)
Reply	0.34 (0.07)	0.16 (0.04)
Quote	0.30 (0.06)	0.19 (0.05)
Response	0.33 (0.07)	0.17 (0.04)

Table 6.4: Random Walk Controversy score for the different simulated user interaction networks.

We calibrate the simulation by setting the *reading density* and the *maximum retweet probability* to values that more closely match the observed empirical average number of retweets (and consequently other interactions) and their relationship to the number of original posts, which are instantiated directly from the empirical distribution. We found that a reading density of 30 (per 4-hour window) and a maximum retweet probability of 18% more closely matched the observed average empirical ratio of Retweets and Original tweets (the former outnumbers the later by a factor of 10) for both types of simulations (with and without the Stance module).

In Table 6.4, we present the average Random Walk Controversy scores obtained for the different simulated interaction networks with and without the inclusion of the Stance module. It is important to note that these results correspond to the final calibrated parameters described before. As we can see, the current implementation of the Twitter Interaction model (without the Stance Module) is not able to recreate the observed level of polarization of the Retweet network. Moreover, we see that this model mostly recreates the low levels of polarization observed in the Follower network, which is consistent with the way that each user's timeline is constructed (see Section 6.3.2 for more details). Importantly, the mechanisms proposed with the stance module are able to recreate both the significant levels of polarization observed in the retweet network and the low levels observed in the responses. However, we see that the polarization levels observed are significantly above the ones observed in the empirical data, specially in the case of the Reply network. The reason behind these possible differences are discussed in the limitations section. In Table 6.5 we present a summary of the main stylized facts observed for the empirical networks, and the ability of the different simulations to recreate them. As we see, the current implementation of both simulations is not able to recreate the different behaviors observed when users reply and quote other users.

	Power Law for User Activity	Non-significant Polarization of Responses	Polarization of Retweets	Different Behavior for Replies and Quotes
Twitter Interaction Model	✓	✓	x	x
Stance Twitter Interaction Model	✓	✓	✓	x

Table 6.5: Summary of Stylized Facts

6.5 Limitations

There are several limitations to the current version of the proposed simulation. First, most of the parameters controlling the probability of interaction of different users were estimated empirically based on observed correlations of the engagement between users of the same and different stances. However, ideally these parameters should be estimated via experimental or quasi-experimental methods, specially when dealing with experiments that will change the type of content a user is exposed to (this was not done in this work). Moreover, the Friends network was crawled more than a year after the studied events transpired and does not fully reflect the true network at the time. For this reason it is not possible to ascertain if the higher levels of polarization observed in the simulated networks is a result of an increased level of polarization in the collected network (with respect to the network at the time) or if it is a function of the mechanism proposed. Finally, the current implementation of the simulation is not able to recreate the different behavior shown by users when deciding whether to Reply or Quote another user. As we have shown throughout this thesis, these interactions can serve different purposes and also be effective in different ways as a response to disinformation.

6.6 Conclusions

In this work we present a case study focused on the "cacerolazo" (pot and pan banging) protests that took place during government-mandated curfews in the height of the 2019 Ecuadorian protests. By leveraging different classifiers of the target and conversation stances of the users during the event, we are able to show that users from both sides of the ideological spectrum actively engaged with each other. Moreover, we show that the polarization levels observed are not a result of filter bubbles and that a user's tendency to mostly share content from others with similar views (via retweets) is more consistent with confirmation bias. Finally, we propose a new Stance module to an existing and validated Twitter Simulation model based on CONSTRUCT's model for information diffusion. This new module augments the agent's cognitive model to reflect their tendency for polarized content engagement. We show, through virtual experiments, that the empirical polarization observed in the Follower network is not *sufficient* to recreate the level of polarization observed in the Retweet network, while our proposed method is able to achieve this goal.

It is important to stress the effect that the data collection methodology applied has on providing a full picture of the nature of the polarization observed during these events. For instance, the common term or hashtag matching methodology is likely to provide a sparse view of the response network (not so for the retweet network) and will likely underestimate the level of connectivity between these groups.

Chapter 7

Conclusions, Limitations and Future Work

7.1 Conclusions

This dissertation aims to address several challenges involved in leveraging user stances in social media conversations to characterize polarized communities and their response to rumorous information during controversial events. The main issues tackled, and the implications of the results obtained, are discussed next.

How does contentious and rumorous information spread through these communities? What effect does the response to this information have on its diffusion? As one of the motivating case studies presented in this thesis, in Chapter 2 we explored different characteristics of disinformation campaigns around the release of two Marvel Blockbusters that experienced significant amounts of contentious discussion. We observed that debunking and mocking quote responses appear to diffuse faster in the community attacking false stories than the false stories themselves. Satirical responses appear to have longer lifetimes and, in some cases, higher speed of diffusion than other false stories. Moreover, our results strongly support the idea that the positive framing around one of these disinformation campaigns motivated participation of influential users. It is noteworthy that, in both case studies, the spike in negative responses coincided with the end of the majority of the support for each disinformation campaign. The purpose of this work should not be taken to be to help in the design of successful misinformation but rather to assist in the understanding of different methods intentionally used to promote false information. This may help with the design of effective and efficient community level interventions.
The above case studies required the identification of known instances of disinformation and the characterization of their responses, a rumor-stance detection problem. In order to generalize this approach to different social contexts and languages, in Chapter 5, we annotated a sample of 7.4 thousand target-response pairs collected around the South American protests. This is one of the largest datasets for this task and the first available in Spanish. We separate the task of detecting stance in conversations in two parts: first we identify agreements, disagreements, or neutral responses; and then we identify whether non-neutral responses are supporting (denying) the veracity of their target. We believe that this unified treatment can help bridge the two related areas of rumor and controversy detection and hope that this work can help inform future research and

decision making regarding the response to false information online.

Can we develop methodologies to identify weak signals in Twitter posts to train large-scale stance detection models in new languages? The usability of existing resources for stance detection can be challenged by the fast-paced way in which content is generated in Social Media. In the context of polarizing events, new important issues emerge suddenly and the way they are discussed can change in a rapid fashion. These issues are exacerbated in languages other than English, as existing resources are scarcer. For this reason it is important to develop stance mining methods that require minimal supervision. In Chapter 3 we proposed a method of the sort, tailored for Twitter conversations during contentious events and apply it to the 2019 South American protests. The weak-labeling approach leverages users' endorsement of politicians' tweets and hashtag campaigns with consistent stances towards the protest (for or against). The focus of the stance annotation effort to a small subset of political figures and hashtag campaigns can drastically reduce the amount of supervision required to produce large-scale semi-supervised datasets. Moreover, the reliance on both signals provides different avenues to assess the robustness of the labels obtained and ameliorate the problem of hashtag hijacking. The mined stances are used to segregate the user pool into two groups: one in favor of the government, and the other, against it. We believe that this methodology holds promise for the development of large-scale databases for the analysis of similar contentious events (with the active involvement of local political figures).

What effect do the different context levels available, as users engage in social media, have on our capacity for stance detection? Users engage in Social Media in a variety of ways, they can post multiple original messages or engage with other users by replying to them or sharing each other's posts. It has long been recognized in the stance detection field that considering only an isolated sentence will provide an incomplete assessment of a user's stance towards a predefined target. However, few research has quantified how leveraging the different context levels available for a given user can affect an algorithm's capacity to infer their stance. In Chapter 4 we explored the value of context in the task of target-stance classification during the protests by using the large-scale weakly-labeled dataset constructed in Chapter 3. The regional nature of the dataset provides the unique opportunity of not only testing the effect of context on in-sample predictions, but to also evaluate the effect on the generalization capabilities of our proposed model in out-of-sample country and temporal data. For this purpose we constructed a compartmentalized architecture that relied on Transformers for the Tweet and User level contexts, and Graph Neural Networks (GNNs) to leverage social media relations. We found that increasing context not only improved the performance of a classifier within the country it was trained, but also made it more robust to out-of-sample predictions. These out-of-sample improvements were substantial both when comparing a classifier's performance across varying country contexts and over time. To the best of our knowledge, the evaluation of the effects of context on the generalization capabilities of these models has not been explored in other work, and has been identified as an outstanding issue in the task of stance classification.

What empirical regularities characterize polarized communities during controversial events? On Chapter 3, we explored the polarization observed throughout the protests both in language and in news media sharing patterns showing that, together, the analyses shed vital insights. Our linguistic polarization results indicate that it largely manifests along ideological, political and protest-related lines; as terms related to extreme positions in left-leaning ideologies in one community are discussed in similar contexts as extreme right-leaning terms in the other. Similarly, we find that opposition leaders are discussed in similar context as government representatives. We also find strong evidence of polarization in the news sharing patterns of users, consistent with their protest stances. This can have pervasive effects on public discourse and political literacy. As a case in point, the results also highlight the important role that regional Russian and Venezuelan news outlets like RT en Español and TeleSUR, played in the social media discussion of the protests and show how effective these outlets have been in gathering an audience of left-leaning users in the region. To the best of our knowledge, our work is one of the first few that combines network-based methods and language-based methods to jointly explore the nature of polarization in a user's news sharing behavior and her language usage.
However, focusing solely on user sharing patterns (based on retweets) is insufficient to understand whether the observed polarization is a function of the lack of access to content from opposing views (due to filter bubbles) or if it is a manifestation of a user's confirmation bias. In Chapter 6 we tackle this issue in the context of the Ecuadorian *Cacerolazo* that occurred during the height of those protests. We leverage the classifiers proposed in Chapter 4 and 5 to determine the protest stance of users during the event and classify the conversation stance of their interactions. We find that even though the Retweet user network is highly polarized, others like the Friend or Response networks, show little or no polarization. Meaning that users from both sides of the ideological spectrum actively engaged with each other during the event. This suggests that the observed polarization is not a result of filter bubbles, and their tendency to mostly retweet content from users with the same stance is more consistent of social media practices exhibiting polarized content engagement (confirmation bias). In this Chapter we also propose a dynamic network simulation model that builds upon an existing Twitter simulation based on CONSTRUCT's model of information diffusion. By using a sample of the empirical Friend network of users in the event, we show that the current implementation of the simulation is unable to recreate the disparate polarization levels observed in the Retweet and Response networks. To account for this we introduce user mechanisms that simulate confirmation bias and are able to recreate the observed levels of polarization in these networks. These results are salient, as policy interventions that focus on changing the way that content is shown to give access to opposing views, without any other type of input or moderation, are not likely to be an effective method for reducing polarization.

The work conducted in this thesis stresses the importance of the data collection methodology on the ability of subsequent research to assess the nature of the polarization observed during these events. For instance, the common term or hashtag matching methodology, often applied in Twitter studies, will only provide a sparse view of the response network (not so for the retweet network). This will likely underestimate the level of connectivity between these groups and hence result in unreliable measures of polarization. Throughout this work we made a considerable effort to rehydrate as much of the different conversation trees as possible during and after each collection.

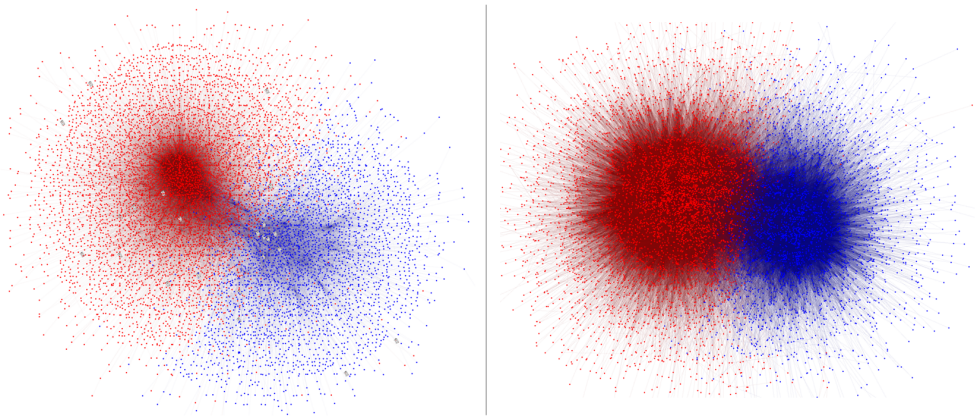

Figure 7.1: Retweet networks for the users that participated in the *cacerolazo* event. The left panel is limited to retweets that occurred during the two days of the event (including the 'cacerola' substring), while the right panel corresponds to all retweets, between the same users, that were part of the Ecuadorian protest collection (approximately 20 days). Users nodes are colored based on the communities determined via graph partitioning, but with labels/colors that align with the predicted protest stance.

7.1.1 Similarities of Protest Stance and Community Detection

Given the importance of the study of polarization in online social media, several frameworks have been proposed for the automatic detection of controversial events. As discussed in this dissertation, the most recent instances generally focus on the user retweet network. One of the most influential of these approaches was proposed by Garimella et al. [46] and allows for the unsupervised detection of controversial events by leveraging the community structure present in the retweet network of a conversation. The unsupervised nature of this method has the advantage of requiring no knowledge of the event in question (other than the set of hashtags that define the event). We could apply this methodology to the *cacerolazo* event, discussed in Chapter 6, by partitioning the user retweet graph in exactly two communities (using an off-the-shelf graph partitioning algorithm) and evaluating the Random Walk Controversy (R.W.C.) score observed for the event. In the left side of Figure 7.1, we present the obtained partitions for the event using METIS [64], which result in a R.W.C. score of 0.70 and would lead us to conclude that we are dealing with a controversial event. This score is not only close to the value we derived with the stance based communities (0.71), but these partitions also align with the predicted protest stances for 96.3% of the users.

The similarity of the labels derived with both approaches is unsurprising, considering that the weak-labeling approach proposed in Chapter 3 explicitly leverages retweets of both political figures and of "stance-tags". However, as also noted in this chapter, the quality of the derived weak-labels is significantly influenced by the exclusivity of the included hashtags. What's more, the inclusion of two different signals allowed us to improve the robustness of the derived labels by consolidating the two results and excluding hashtags that were too generally used by the two groups. To see this, consider the retweet network between the same set of users identified before,

but now including all their retweets within the Ecuadorian collection. In the right panel of Figure 7.1, we present the results of the described methodology in this new setting. We note that the new partitions have lower modularity, their alignment with the underlying stance labels is significantly lower than before (87.6% of users match) and there is also a significant decrease in the R.W.C. score for the conversation (0.57). The latter result underlines that, as discussed throughout this thesis, the focus on the retweet network during a specific event can greatly underestimate the exposure that users with different stances have to each other's content. Furthermore, the alignment between the two partitions and the underlying stance of the users will continue to degrade as we consider the full set of users that where part of the Ecuadorian protests collection. This will occur as the inclusion of different conversations regarding specific events (even within the context of the protest) will result on a retweet graph that is less clearly partitioned between two groups.

The final important distinction comes from our ability to extrapolate the results from both methodologies. On could try to reduce the disparity between the partitions and the stance of the users by focusing only in the retweet network of "stance-tags" (or a subset therein). However, this will also greatly reduce the number of users considered, as it will exclude users that do not retweet the selected hashtags. For example, our weak-labeling methodology is only able to assign a label to approximately 5% of the users included in the collection. The classifiers developed in Chapter 4 are able to undercut this limitation by leveraging both text and interaction features. In this way we are able to assign a stance to users that never used "stance-tags" or that never retweeted another user. Moreover, it allowed us to show how leveraging these different contexts can improve the generalization capabilities of these models. Importantly, the prediction of the Chilean referendum vote using the models trained with data a year prior, would not have been possible by applying community detection on the retweet network of the protests data as 46% of the users were not seen in the original collection.

7.2 Limitations

In this section I list the main limitations of the research presented in this thesis. First, the proposed weak-labeling methodology is not able to determine users with neutral stances towards the government (or the protests). This limits the scope of the analysis presented in this work to users with two defined stances towards the event: for or against. It also implies that the protest stance classifier developed may incorrectly categorize users with no or contradictory/nuanced opinions into one of the two categories. Secondly, the different proposed neural architectures only use interactions (reply and quote, retweets) and the text of these users for classification. This excludes many other attributes available on Twitter that might correlate with stance, including URLs, followership, posting patterns, bios, and shared multimedia. Moreover, our the different conversation stance classifiers are trained on tweet pairs that occur in the first level of the conversation tree. This can degrade the classification performance of the models on instances that occur further below the tree, as the context of the interaction is less dependent on what is observed in the pair.

In addition, we did not attempt to remove bots or trolls from the datasets, nor did we survey in-country Twitter user demographics, so these datasets may be noisy proxies for population

opinions. However, despite these limitations we found the predicted stance distribution of users in the Chilean referendum data was nearly identical to the final referendum vote. With regards to our study of the *cacerolazo* event, discussed in Chapter 6, we can test the robustness of the results to the inclusion of bots with the usage of an off-the-shelf bot detection model [**?**]. Even though the application of this algorithm results in the removal of 23.1% of users for the event, the R.W.C. score for the user retweet network increases to 0.75 (a 0.04 increase) and is unchanged for the response network. This implies that the results obtained in that chapter are robust to the inclusion of bots in the conversations studied.
Finally, given the difficulty of assessing the truthfulness of Tweets, this work focuses mainly on rumorous tweets. The only requirement for a tweet to be rumorous is that responses to it accuse it of containing false information. Hence, the results obtained are limited to user perceptions, and not a true assessment of whether the tweet is true or false. Moreover, we only focus on direct interactions on Twitter as indicators of exposure. This ignores Likes (not available on the free API) or tweets seen by users but where no interaction occurs. Similarly, we limit our analysis to Twitter, and do not explore how the phenomena described manifest in other social media or is reinforced by it. In what follows we will review the limits in generalization and scalability of the proposed algorithms and resources.

7.2.1 Generalization of Proposed Algorithms

It is important to underline the generalization limits of the classifiers and resources proposed in this dissertation and the scope of conditions where they can be applied with or without fine-tuning.

Weak-labeling methodology The methodology proposed in Chapter 3 holds promise for the development of large-scale databases for the analysis of stance during contentious events. However, the reliance on the endorsement of political figures as a signal, though important to improve the robustness of the derived labels, limits the scope of this methodology to highly political events with active involvement of local political figures.

twBETO v0 and v1 In Chapter 4 and 5 we presented two BERT language models specialized for Spanish Twitter conversations, with the latter improving on several key aspects of the former. As we showed in Chapter 5, this model significantly outperforms other Spanish language models for Twitter, like *TwilBERT* [62], as it has a better coverage of non-European Spanish variants, is trained on more data and applies the RoBERTa pretraining framework. However, this model was trained in the context of Twitter conversations, and care should be taken when applying it to Spanish text from other sources as performance is likely to degrade [80, 94].

Protest Stance Classifiers In Chapter 4 we proposed several classifiers for the stance of users during the 2019 South American protests. We also showed that, by leveraging the different levels of context available, the heterogeneous GNN model is able to generalize better to different country-contexts and to future data. Importantly, we also stressed the limits of this generalization

capability with the Bolivian case, as the diametrically opposed motivations for the protests served as an adversarial domain for these classifiers. Moreover, it is unlikely that finetuning the other country models on the Bolivian data, instead of the one-shot methodology applied, would be able to address this issue. This highlights the importance of domain knowledge when applying these classifiers, and deep learning models in general, to different domains.

7.2.2 Scalability of Proposed Algorithms

One common limitation of research that utilizes Deep Learning architectures consisting of tens or hundreds of millions of parameters, is the scalability of the models to the size of the data often encountered in real social networks. In order to make these models somewhat viable for big data analysis, it is often necessary to have access to specialized hardware consisting of one or more GPUs with considerable amounts of memory. The work undertaken in this thesis is no exception, given the ample usage of Transformer-based architectures. In Table 7.1 we provide a brief description of the computational cost (in hours) of training these models or using them for inference. As we can see, most of the Transformer-based architectures (all but the GNN) are somewhat scalable, only if there is access to specialized GPUs with more than 16 GBs of RAM. The least efficient of these models is the User Level protest stance classifier as it stacks two different Transformer models. The GNN based models are considerably more scalable, but require as input user embeddings that, in this work, are computed by the User-Level classifier.

Chapter	Model	Data Type	Training (h)	Inference (h)	Hardware Used
	twBETO v0	Single or Pair of Tweets	650 (150M tw)	7 (1M p)	Titan XP
4	User-Level (S.C.)	Batch of User Tweets	16 (250k u)	6 (250k u)	Titan XP
	Hetero. GNN S.C.	User Emb. + Social Net.	1 (250k u)	0.4 (250k u)	Titan XP
5	twBETO v1	Single or Pair of Tweets	500 (201M tw)	4 (1M p)	RTX A6000

Table 7.1: Computational cost (in hours) of the different classifiers proposed based on the GPU used. The different type of inputs for a model vary from single tweet (tw), pair of tweets (p), or user (u).

7.3 Ethical Considerations

In the pursuit of the objectives set forth in this work we labeled, either manually or through weak supervision, the stance of thousands of tweets and users. For the most part, only the IDs of the tweets are shared, ensuring that tweets will have to be re-hydrated through Twitter's API. However, to preserve the usability of the annotated conversation stance dataset over time, it was necessary to also share an anonymized version of the text **only for the annotated pairs**. This was obtained by by masking any user mentions or URLs contained in the tweet. However, to adhere to Twitter's terms and conditions for sharing data, we do not share any user IDs or information. To obtain this information a tweet will have to be re-hydrated, so if a user deletes a tweet (or their account), their information will no longer be available ensuring that their *right to be forgotten* is preserved. Moreover, given the inclusion of anonymized text data with the annotated labels,

and to adhere to Twitter's fair usage policy, we opt for a CC BY-NC licence for this resource. This dataset was created to provide a unified resource for the joint exploration of rumors and controversy during contentious events. It is intended for research purposes and should not be used for commercial purposes (as indicated by the chosen licence). According to Twitter's fair use policy, research based on this data should be presented at a aggregate level and should not be used for the identification of individual users.
Finally, the purpose of this work should not be taken to be to help in the design of successful misinformation but rather to assist in the understanding of different methods intentionally used to promote false information. This may help with the design of effective and efficient community level interventions – at a minimum as an example of the kinds of campaigns to be aware of and vigilant against. Future work should delve more deeply into the cross-platform and cross-network nature of these conversations with an eye toward how that may improve our ability to classify the intent and effect of various campaigns.

7.4 Future Work

In this section I would like to suggest several future avenues of research that will help to advance the state of the art on the areas explored in this work.

Classification of Users with Neutral Stances. The current implementation of the target-stance classifiers, and the weak-labeling methodology used to train them, excludes the possibility of characterizing users with neutral stances. However, this is an important group and estimating their prevalence in these events merit deeper exploration as they can potentially serve as the bridge between polarized communities. In Chapter 5 we took a first step in this direction, by creating annotated resources identifying neutral stances towards the different governments and protests. However, work needs to be done to integrate these type of stances in semi-supervised methods to produce datasets of the sizes often required for data-hungry statistical models.

Multi-modal and cross platform characterization of polarized communities. The current work was focused primarily on text-based signals revealed by users in the Twitter social media platform. However, richer features, relying on other modalities of data, can be extracted from these sites. As we have shown in Chapter 4, exploiting additional context can be an important way to not only improve the in-sample performance of algorithms, but also to make them more generalizable. In a similar vein, it is also important for future work to explore cross-platform interactions between users to obtain a better assessment of the nature of these polarized communities.

Exploiting User and Network Contexts for Conversation Stance Classification. The dataset released in Chapter 5 contains multiple different aspects of the stance revealed by users during the protests. It also includes the IDs of millions of other tweets that were part of these conversations. However, the models trained in that chapter only focused on a pair of tweets and ignored the other context available. Moreover, as was shown with the more conservative version of the classifiers trained, this dataset provides an important resource for multitask learning. The exploitation of this additional context and the multitask nature of the data, can be a fruitful vein

of research for future work on the area of conversation stance learning.

Causal exploration of rumor-stance detection and confirmation bias. The work undertaken in this thesis focused primarily on observational data and the analysis of correlation. Work remains to be done to explore the causal effects of polarization in discussions and how, for example, it can affect the existing distrust between users. There are two main ways to approach this problem. The first is by the design of controlled experiments to quantify this distrust, and the second through the causal analysis of observational data (quasi-causal methods). For an example of the latter, define the treatment as being exposed to a message from a user with a different stance and the outcome as whether a user then denies (or affirms) the veracity of the read message. We can then leverage the classifiers built in Chapter 5 to identify response instances where users deny (affirm) the veracity of their target and the classifier in Chapter 4 to identify when those interactions correspond to users with the same or different stances. To recover a causal estimation of this confirmation bias effect one could, for example, apply matching methods (propensity score matching, coarsened exact matching, etc.) to algorithmically identify "similar" instances based on covariates like the homophily in the users' social network and the semantic content of the message. Similarly, one can leverage the temporal structure present in the data to identify the effect of debunking responses in their target. We explored this question in Chapter 2 where, in the context of the Black Panther case study, we identified that users subject to debunking responses were less likely to continue to spread the false stories. Work remains to be done to identify if this effect is also observed in other, more political events, and to estimate the effect that the possible difference in the stance of users can have on the effectiveness of debunking responses.

Continuous Exploration of the South American Political Environment. The repercussions of the 2019 South American Protests were not only felt that year, but have continued to reshape the political landscape of the region. Protests continued through the Coronavirus pandemic in all these countries and, in the case of Chile, led to a new Constitution and the election of a new government largely supportive of the original protests. In Colombia, massive protests took place in 2021 with far more reported violence, both on the side of government and protesters, than what was observed in 2019. As a result of all these changes, on June of 2022, Colombia elected its first left-wing government in history. Similarly, at the time of this writing, massive protests largely reminiscent of 2019, are underway in Ecuador. The continual exploration of the ongoing political instability in the region provides a unique opportunity for longitudinal analysis of the evolution of the protests and the recorded levels of polarization in language and sharing patterns. This can serve as an important resource to address the always present question in this type of studies, namely, *How does Social Media relate to the real world?* In Chapter 4 we took a stab at this question by showing that the protest stance of users during the Chilean protests was an effective predictor of the vote for the constitutional referendum that took place a year later. However, through a longitudinal study of the different iterations of these protests one could quantify the effect, if any, that the polarization observed in social media has had in the trust of local media outlets in each of these countries, and the reported levels of violence derived from the protests.

Bibliography

[1] $6.78 billion public opinion and election polling global market to 2030 - identify growth segments for investment, Jul 29 2021. URL https://www.proquest.com/wire-feeds/6-78-billion-public-opinion-election-polling/docview/2555898927/se-2?accountid=9902. Copyright - GlobeNewswire, Inc; Last updated - 2021-07-29. 4.1

[2] Mahmoud Al-Ayyoub, Abdullateef Rabab'ah, Yaser Jararweh, Mohammed N Al-Kabi, and Brij B Gupta. Studying the controversy in online crowds' interactions. *Applied Soft Computing*, 66:557–563, 2018. 1.1, 5.1

[3] Abdulrahman I Al-Ghadir, Aqil M Azmi, and Amir Hussain. A novel approach to stance detection in social media tweets by fusing ranked lists and sentiments. *Information Fusion*, 67:29–40, 2021. 1.1, 4.2

[4] Khaled Alhazmi, Walaa Alsumari, Indrek Seppo, Lara Podkuiko, and Martin Simon. Effects of annotation quality on model performance. In *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 063–067, 2021. 3.1

[5] Giambattista Amati, Simone Angelini, Giorgio Gambosi, Daniele Pasquin, Gianluca Rossi, and Paola Vocca. Twitter: temporal events analysis. In *Proceedings of the 4th EAI International Conference on Smart Objects and Technologies for Social Good*, pages 298–303. ACM, 2018. 1

[6] Lampinen Andrew K., Dasgupta Ishita, Chan Stephanie C. Y., Matthewson Kory, Tessler Michael Henry, Creswell Antonia, McClelland James L., and Hill Jane X., Wang & Felix. Can language models learn from explanations in context? *DeepMind*, 2022. 5.2

[7] Ahmer Arif, John J Robinson, Stephanie A Stanek, Elodie S Fichet, Paul Townsend, Zena Worku, and Kate Starbird. A closer look at the self-correcting crowd: Examining corrections in online rumors. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 155–168, 2017. 1, 6.1

[8] Isabelle Augenstein, Andreas Vlachos, and Kalina Bontcheva. 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)*, pages 389–393, 2016. 5.2

[9] Matthew Babcock, David M Beskow, and Kathleen M Carley. Beaten up on twitter? exploring fake news and satirical responses during the black panther movie event. In *Inter-*

national Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, pages 97–103. Springer, 2018. 2.4.1

[10] Matthew Babcock, David M Beskow, and Kathleen M Carley. Different faces of false: The spread and curtailment of false information in the black panther twitter discussion. *Journal of Data and Information Quality (JDIQ)*, 11(4):18, 2019. 1, 1.2.1, 2.3

[11] Matthew Babcock, Ramon Villa-Cox Cox, and Sumeet Kumar. Diffusion of pro-and anti-false information tweets: the black panther movie case. *Computational and Mathematical Organization Theory*, 25(1):72–84, 2019. 1, 2.1, 2.4, 3.2, 5.1

[12] Matthew Babcock, Ramon Villa-Cox, and Kathleen M Carley. Pretending positive, pushing false: Comparing captain marvel misinformation campaigns. In *Disinformation, Misinformation, and Fake News in Social Media*, pages 83–94. Springer, 2020. 2.1, 2.4, 3.2

[13] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528, 2012. 6.2.2

[14] Matthew Baldwin and Joris Lammers. Past-focused environmental comparisons promote proenvironmental outcomes for conservatives. *Proceedings of the National Academy of Sciences*, 113(52):14953–14957, 2016. 3.2

[15] Pablo Barberá, Ning Wang, Richard Bonneau, John T Jost, Jonathan Nagler, Joshua Tucker, and Sandra González-Bailón. The critical periphery in the growth of social protests. *PloS one*, 10(11):e0143611, 2015. 3.2

[16] Mariano Beguerisse-Díaz, Guillermo Garduno-Hernández, Borislav Vangelov, Sophia N Yaliraki, and Mauricio Barahona. Interest communities and flow roles in directed networks: the twitter network of the uk riots. *Journal of The Royal Society Interface*, 11 (101):20140940, 2014. 3.2

[17] Alessandro Bessi, Fabiana Zollo, Michela Del Vicario, Michelangelo Puliga, Antonio Scala, Guido Caldarelli, Brian Uzzi, and Walter Quattrociocchi. Users polarization on facebook and youtube. *PloS one*, 11(8):e0159641, 2016. 3.2, 6.4.1

[18] Steven Bird, Ewan Klein, and Edward Loper. Nltk book, 2009. 4.4.1, 5.3.3

[19] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. ISSN 2307-387X. 3.4

[20] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *2010 43rd Hawaii International Conference on System Sciences*, pages 1–10. IEEE, 2010. 1

[21] William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318, 2017. 2.2

[22] Cody Buntain and Jennifer Golbeck. Automatically identifying fake news in popular twitter threads. In *2017 IEEE International Conference on Smart Cloud (SmartCloud)*,

pages 208–215. IEEE, 2017. 1, 1.1, 1.3, 5.1

[23] Lauren M Burch, Evan L Frederick, and Ann Pegoraro. Kissing in the carnage: An examination of framing on twitter during the vancouver riots. *Journal of Broadcasting & Electronic Media*, 59(3):399–415, 2015. 3.2

[24] Paul Burstein. The impact of public opinion on public policy: A review and an agenda. *Political research quarterly*, 56(1):29–40, 2003. 4.1

[25] Kathleen M Carley. Ora: A toolkit for dynamic network analysis and visualization. *Encyclopedia of social network analysis and mining*, pages 1219–1228, 2014. 2.3

[26] Kathleen M Carley, Michael K Martin, and Brian R Hirshman. The etiology of social change. *Topics in Cognitive Science*, 1(4):621–650, 2009. 6.1, 6.2.2

[27] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684, 2011. 2.2, 5.1

[28] Hélène Combis. De la monarchie de juillet à françois fillon, petite histoire de la casserole comme outil politique, Apr 2022. URL `https://www.franceculture.fr/histoire/de-la-monarchie-de-juillet-francois-fillon-petite-histoire-de-la-casse` 6.2.1

[29] Kareem Darwish. Quantifying polarization on twitter: the kavanaugh nomination. *arXiv*, abs/2001.02125, 2020. 3.2

[30] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559, 2016. 3.2

[31] Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In *NAACL-HLT 2019*, pages 2970–3005. Association for Computational Linguistics, 2019. 3.2

[32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 1.3, 4.1, 4.4.1, 5.2

[33] Taylor Dewey, Juliane Kaden, Miriam Marks, Shun Matsushima, and Beijing Zhu. The impact of social media on social unrest in the arab spring. *International Policy Program*, 5(8):1–76, 2012. 3.2

[34] Paul DiMaggio and Kyoko Sato. Does the internet balkanize political attention?: A test of the sunstein theory. In *Annual meeting of the American Sociological Association, Atlanta*, 2003. 3.2

[35] Stephen Dipple, Michael Kowalchuck, Neal Altman, and Kathleen M Carley. Construct user guide. 2022. 6.1, 6.3.2, **??**, **??**

[36] Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. Stance classification with target-specific

neural attention networks. International Joint Conferences on Artificial Intelligence, 2017. 4.2

[37] John W Du Bois. The stance triangle. *Stancetaking in discourse: Subjectivity, evaluation, interaction*, 164(3):139–182, 2007. 1.1, 4.1, 4.2

[38] Hillard Dustin and Purpura & John Wilkerson Stephen. Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics*, 2008. 5.2

[39] Omar Enayet and Samhaa R El-Beltagy. Niletmrg at semeval-2017 task 8: Determining rumour and veracity support for rumours on twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 470–474, 2017. 5.1

[40] Ash Evans. Stance and identity in twitter hashtags. *Language@ internet*, 13(1), 2016. 3.2, 3.3.1

[41] Emilio Ferrara. Manipulation and abuse on social media by emilio ferrara with ching-man au yeung as coordinator. *ACM SIGWEB Newsletter*, (Spring):4, 2015. 2.2

[42] Dana R. Fisher, Joseph Waggle, and Philip Leifeld. Where does political polarization come from? locating polarization within the us climate change debate. *American Behavioral Scientist*, 57(1):70–92, 2013. 3.2

[43] Seth Flaxman, Sharad Goel, and Justin M Rao. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298–320, 2016. 3.2

[44] Ryan J Gallagher, Andrew J Reagan, Christopher M Danforth, and Peter Sheridan Dodds. Divergent discourse between protests and counter-protests:# blacklivesmatter and# alllivesmatter. *PloS one*, 13(4):e0195644, 2018. 3.2

[45] Kiran Garimella, Ingmar Weber, and Munmun De Choudhury. Quote rts on twitter: usage of the new feature for political discourse. In *Proceedings of the 8th ACM Conference on Web Science*, pages 200–204. ACM, 2016. 5.1, 6.2.2

[46] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):1–27, 2018. 5.1, 5.3.1, 6.3.1, 6.4.1, 7.1.1

[47] R Kelly Garrett. Politically motivated reinforcement seeking: Reframing the selective exposure debate. *Journal of communication*, 59(4):676–699, 2009. 6.4.1

[48] R. Kelly Garrett. The "echo chamber" distraction: Disinformation campaigns are the problem, not audience fragmentation. *Journal of Applied Research in Memory and Cognition*, 6(4):370 – 376, 2017. ISSN 2211-3681. URL `http://www.sciencedirect.com/science/article/pii/S2211368117301936`. 3.2

[49] Matthew Gentzkow and Jesse M Shapiro. Ideological segregation online and offline. *The Quarterly Journal of Economics*, 126(4):1799–1839, 2011. 3.2

[50] Jennifer Golbeck and Derek Hansen. Computing political preference among twitter followers. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1105–1108, 2011. 3.2

[51] Sandra González-Bailón and Ning Wang. Networked discontent: The anatomy of protest campaigns in social media. *Social networks*, 44:95–104, 2016. 3.2

[52] Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. Semeval-2019 task 7: Rumoureval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, 2019. 5.1

[53] Yupeng Gu, Ting Chen, Yizhou Sun, and Bingyu Wang. Ideology detection for twitter users via link analysis. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 262–268. Springer, 2017. 3.2

[54] Julia Gurganus. Russia: Playing a geopolitical game in latin america. *Carnegie Endownent for Peace*, 2018. 3.5.1, 3.7

[55] William L Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1025–1035, 2017. 4.4.3

[56] Dirk Hovy and Shannon L Spruit. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, 2016. 3.1

[57] Binxuan Huang. *Learning User Latent Attributes on Social Media*. PhD thesis, Carnegie Mellon University, 2019. 2.3

[58] Keisuke Ikeda, Yoshiyuki Okada, Fujio Toriumi, Takeshi Sakaki, Kazuhiro Kazama, Itsuki Noda, Kosuke Shinoda, Hirohiko Suwa, and Satoshi Kurihara. Multi-agent information diffusion model for twitter. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 21–26. IEEE, 2014. 6.2.2

[59] Tiddi Ilaria and Schlobach Stefan. Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence*, 302, 2022. 5.2

[60] Lee J., Yoon W., Kim S., Kim D., Kim S., and So & J. Kang C.H. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics)*, 36(4):1234–, 2020. 5.2

[61] Kenneth Joseph, Lisa Friedland, William Hobbs, Oren Tsur, and David Lazer. Constance: Modeling annotation contexts to improve stance classification. *arXiv preprint arXiv:1708.06309*, 2017. 1.1

[62] González José Ángel and Pla Lluís-F., Hurtado & Ferran. Twilbert: Pre-trained deep bidirectional transformers for spanish twitter. *Neurocomputing)*, 426:58–69, 2020. 4.1, 5.2, 5.4.2, **??**, **??**, 7.2.1

[63] Márton Karsai, Mikko Kivelä, Raj Kumar Pan, Kimmo Kaski, János Kertész, A-L Barabási, and Jari Saramäki. Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E*, 83(2):025102, 2011. 2.2

[64] George Karypis and Vipin Kumar. A software package for partitioning unstructured

graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices. *University of Minnesota, Department of Computer Science and Engineering, Army HPC Research Center, Minneapolis, MN*, 38, 1998. 7.1.1

[65] Ashiqur R KhudaBukhsh, Rupak Sarkar, Mark S Kamlet, and Tom M Mitchell. Fringe news networks: Dynamics of us news viewership following the 2020 presidential election. *arXiv preprint arXiv:2101.10112*, 2021. 3.2

[66] Ashiqur R. KhudaBukhsh, Rupak Sarkar, Mark S. Kamlet, and Tom M. Mitchell. We don't speak the same language: Interpreting polarization through machine translation. In *AAAI 2021*, page To Appear. AAAI Press, 2021. 3.1, 3.2, 3.4

[67] Ashiqur R. KhudaBukhsh, Rupak Sarkar, Mark S. Kamlet, and Tom Michael Mitchell. Fringe news networks: Dynamics of us news viewership following the 2020 presidential election. *ArXiv*, abs/2101.10112, 2021. 3.1

[68] Yoon Kim. Convolutional neural networks for sentence classification. In *EMNLP 2014*, pages 1746–1751, October 2014. 3.5

[69] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4.4.1, 5.3.3

[70] Maksim Kitsak, Lazaros K Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H Eugene Stanley, and Hernán A Makse. Identification of influential spreaders in complex networks. *Nature physics*, 6(11):888–893, 2010. 2.2

[71] Kevin Knight, Ani Nenkova, and Owen Rambow. Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Human language technologies. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016. 5.1

[72] Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. All-in-one: Multi-task learning for rumour verification. *arXiv preprint arXiv:1806.03713*, 2018. 1.1, 4.2

[73] Danai Koutra, Paul N. Bennett, and Eric Horvitz. Events and controversies: Influences of a shocking news event on information seeking. *CoRR*, abs/1405.1486, 2014. URL `http://arxiv.org/abs/1405.1486`. 3.5.2

[74] David Krackhardt and Robert N Stern. Informal networks and organizational crises: An experimental simulation. *Social psychology quarterly*, pages 123–140, 1988. 6.3.1

[75] Dilek Küçük and Fazli Can. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37, 2020. 1.1, 1.1, 1.3, 4.1, 4.2, 4.2, 5.1, 5.2, 5.4.2

[76] Sumeet Kumar, Ramon Villa Cox, Matthew Babcock, and Kathleen M Carley. A weakly supervised approach for classifying stance in twitter replies. *arXiv preprint arXiv:2103.07098*, 2021. 5.1, 5.3.1, 5.3.2

[77] Mikko Laitinen, Masoud Fatemi, and Jonas Lundberg. Size matters: Digital social networks and language change. *Frontiers in Artificial Intelligence*, 3:46, 2020. 4.5.2

[78] Jeffrey R Lax and Justin H Phillips. The democratic deficit in the states. *American Journal of Political Science*, 56(1):148–166, 2012. 4.1

[79] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018. 1, 6.1

[80] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. 7.2.1

[81] Kristina Lerman and Rumi Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 4, 2010. 6.2.2

[82] Jure Leskovec and Christos Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636, 2006. 6.3.2

[83] Ziyi Li, Junpei Kawamoto, Yaokai Feng, and Kouichi Sakurai. Cyberbullying detection using parent-child relationship between comments. In *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services*, pages 325–334, 2016. 1

[84] Rich Ling. Confirmation bias in the era of mobile news consumption: The social and psychological dimensions. *Digital Journalism*, pages 1–9, 2020. 3.2

[85] Can Liu, Wen Li, Bradford Demarest, Yue Chen, Sara Couture, Daniel Dakota, Nikita Haduong, Noah Kaufman, Andrew Lamont, Manan Pancholi, et al. 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)*, pages 394–400, 2016. 5.2

[86] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 4.4.1, 5.3.3, 5.4.2

[87] Joshi M., Chen D., Liu Y., Weld D.S., and Zettlemoyer & O. Levy L. Spanbert: Improving pre-training by representing and predicting spans. *arXiv:1907.10529v3*, 2019. 5.2

[88] Jing Ma, Wei Gao, and Kam-Fai Wong. Detect rumor and stance jointly by neural multi-task learning. In *Companion proceedings of the the web conference 2018*, pages 585–593, 2018. 1.1

[89] Sofus A Macskassy and Matthew Michelson. Why do people retweet? anti-homophily wins the day! In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011. 1

[90] C. McConnell, Y. Margalit, N. Malhotra, and M Levendusky. Research: Political polarization is changing how americans work and shop. *Harvard Business Review*, 2017. 3.2

[91] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41, 2016. 1.1, 1.1, 3.1, 3.2

[92] Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):26, 2017. 1.1, 4.2, 5.2

[93] Rachel Mourao and Weiyue Chen. Covering protests on twitter: The influences on journalists' social media portrayals of left- and right-leaning demonstrations in brazil. *The International Journal of Press/Politics*, 25, 10 2019. doi: 10.1177/1940161219882653. 3.2

[94] Dat Quoc Nguyen, Thanh Vu, and Anh-Tuan Nguyen. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, 2020. 4.4.1, 7.2.1

[95] Yoshiyuki Okada, Keisuke Ikeda, Kosuke Shinoda, Fujio Toriumi, Takeshi Sakaki, Kazuhiro Kazama, Masayuki Numao, Itsuki Noda, and Satoshi Kurihara. Sir-extended information diffusion model of false rumor and its prevention strategy for twitter. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 18(4):598–607, 2014. 6.2.2

[96] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13, 2019. 3.1

[97] Keith T Poole and Howard. The polarization of american politics. *The journal of politics*, 46(4):1061– 1079, 1984. 3.2

[98] Keith T Poole and Howard Rosenthal. The polarization of american politics. *The journal of politics*, 46(4):1061–1079, 1984. 3.1

[99] Ana-Maria Popescu and Marco Pennacchiotti. Detecting controversial events from twitter. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1873–1876, 2010. 1.1

[100] Markus Prior. Media and political polarization. *Annual Review of Political Science*, 16: 101–127, 2013. 3.2

[101] Rob Procter, Farida Vis, and Alex Voss. Reading the riots on twitter: methodological innovation for the analysis of big data. *International journal of social research methodology*, 16(3):197–214, 2013. 5.1

[102] William Proctor and Bridget Kies. On toxic fan practices and the new culture wars. *Participations*, 15(1):127–142, 2018. 2.1

[103] Manoel Horta Ribeiro, Pedro H Calais, Virgílio AF Almeida, and Wagner Meira Jr. " everything i disagree with is# fakenews": Correlating political polarization and spread of misinformation. *arXiv preprint arXiv:1706.05924*, 2017. 2.1

[104] Vladimir Rouvinski. Understanding russian priorities in latin america. *Kennan Cable*, 20, 2017. 3.5.1, 3.7

[105] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 4.4.1

[106] Arno Scharl, Albert Weichselbraun, and Wei Liu. Tracking and modelling information

diffusion across interactive online media. *International Journal of Metadata, Semantics and Ontologies*, 2(2):135–145, 2007. 1

[107] Frank Schweitzer and David Garcia. An agent-based model of collective emotions in online communities. *The European Physical Journal B*, 77(4):533–545, 2010. 6.2.2

[108] Umme Aymun Siddiqua, Abu Nowshed Chy, and Masaki Aono. 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)*, pages 1868–1873, 2019. 1.1, 4.2

[109] Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations, ICLR 2017*, 2017. 3.4

[110] Pawel Sobkowicz, Michael Kaschesky, and Guillaume Bouchard. Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. *Government information quarterly*, 29(4):470–479, 2012. 6.2.2

[111] Swapna Somasundaran and Janyce Wiebe. 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*, pages 116–124. Association for Computational Linguistics, 2010. 5.2

[112] Dominic Spohr. Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review*, 34(3):150–160, 2017. 3.2, 6.4.1

[113] Stefan Stieglitz and Linh Dang-Xuan. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of management information systems*, 29(4):217–248, 2013. 2.2

[114] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *2010 IEEE Second International Conference on Social Computing*, pages 177–184. IEEE, 2010. 2.2

[115] Sandesh Swamy, Alan Ritter, and Marie-Catherine de Marneffe. "i have a feeling trump will win.....": Forecasting winners and losers from user predictions on twitter. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1583–1592, 2017. 3.2

[116] Edson C Tandoc Jr, Zheng Wei Lim, and Richard Ling. Defining "fake news" a typology of scholarly definitions. *Digital journalism*, 6(2):137–153, 2018. 1, 2.2

[117] Mariona Taulé, M Antonia Martí, Francisco M Rangel, Paolo Rosso, Cristina Bosco, Viviana Patti, et al. 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, pages 157–177. CEUR-WS, 2017. 5.1, 5.2

[118] Mariona Taulé, Francisco M Rangel Pardo, M Antònia Martí, and Paolo Rosso. Overview of the task on multimodal stance detection in tweets on catalan# 1oct referendum. In

IberEval@ SEPLN, pages 149–166, 2018. 3.2

[119] Rudra M Tripathy, Amitabha Bagchi, and Sameep Mehta. A study of rumor control strategies on social networks. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1817–1820, 2010. 6.1

[120] Adam Tsakalidis, Nikolaos Aletras, Alexandra I Cristea, and Maria Liakata. Nowcasting the stance of social media users in a sudden vote: The case of the greek referendum. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 367–376, 2018. 3.2, 4.1

[121] Sebastián Valenzuela, Nicolás M Somma, Andrés Scherman, and Arturo Arriagada. Social media in latin america: deepening or bridging gaps in protest participation? *Online information review*, 2016. 3.2

[122] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4.4.2

[123] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *Proceedings of the Sixth International Conference on Learning Representations*, 2018. 4.4.3

[124] Michela Del Vicario, Walter Quattrociocchi, Antonio Scala, and Fabiana Zollo. Polarization and fake news: Early warning of potential misinformation targets. *ACM Transactions on the Web (TWEB)*, 13(2):1–22, 2019. 2.2

[125] Ramon Villa-Cox, Ashiqur R KhudaBukhsh, Kathleen M Carley, et al. Exploring polarization of users behavior on twitter during the 2019 south american protests. *arXiv preprint arXiv:2104.05611*, 2021. 4.1, 4.5.1, 5.4.1

[126] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018. 1, 2.2, 6.1

[127] Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. 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)*, pages 384–388, 2016. 5.2

[128] Lilian Weng, Alessandro Flammini, Alessandro Vespignani, and Fillipo Menczer. Competition among memes in a world with limited attention. *Scientific reports*, 2:335, 2012. 6.2.2, 6.4.2

[129] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019. 4.4.1, 5.3.3

[130] Jonas Wolff. Amérique latine : qu'ont obtenu les mobilisations populaires? *Alternatives Economiques*, (122):44–45, 2021. URL `https://blog.prif.org/2020/12/17/one-year-later-the-legacy-of-latin-americas-2019-mass-protests/`. 1, 5.1

[131] Gadi Wolfsfeld, Elad Segev, and Tamir Sheafer. Social media and the arab spring: Politics comes first. *The International Journal of Press/Politics*, 18(2):115–137, 2013. 3.2

[132] Felix Ming Fai Wong, Chee Wei Tan, Soumya Sen, and Mung Chiang. Quantifying political leaning from tweets, retweets, and retweeters. *IEEE transactions on knowledge and data engineering*, 28(8):2158–2172, 2016. 3.2

[133] Zhiyuan Wu, Dechang Pi, Junfu Chen, Meng Xie, and Jianjun Cao. Rumor detection based on propagation graph neural network with attention mechanism. *Expert systems with applications*, 158:113595, 2020. 4.2

[134] Zhiping Xiao, Weiping Song, Haoyan Xu, Zhicheng Ren, and Yizhou Sun. Timme: Twitter ideology-detection via multi-task multi-relational embedding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2258–2268, 2020. 3.2

[135] Fei Xiong, Yun Liu, Zhen-jiang Zhang, Jiang Zhu, and Ying Zhang. An information diffusion model based on retweeting mechanism for online social media. *Physics Letters A*, 376(30-31):2103–2108, 2012. 2.2

[136] Ruifeng Xu, Yu Zhou, Dongyin Wu, Lin Gui, Jiachen Du, and Yun Xue. Overview of nlpcc shared task 4: Stance detection in chinese microblogs. In *Natural language understanding and intelligent applications*, pages 907–916. Springer, 2016. 5.1

[137] Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., and Stoyanov L., Zettlemoyer & V. Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692v1*, 2019. 5.2

[138] Jaewon Yang and Jure Leskovec. Modeling information diffusion in implicit networks. In *2010 IEEE International Conference on Data Mining*, pages 599–608. IEEE, 2010. 2.2

[139] Jiang Yang and Scott Counts. Predicting the speed, scale, and range of information diffusion in twitter. In *Fourth International AAAI Conference on Weblogs and Social Media*. Citeseer, 2010. 2.2

[140] Lan Zhenzhong, Chen Mingda, Goodman Sebastian, Gimpel Kevin, and Soricut Piyush, Sharma & Radu. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020. 5.2

[141] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Kalina Bontcheva, and Peter Tolmie. Crowdsourcing the annotation of rumourous conversations in social media. In *Proceedings of the 24th International Conference on World Wide Web*, pages 347–353, 2015. 1, 1.1, 4.2, 5.1, 5.2, 5.3.2, 5.4.1, 5.4.1

[142] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Kalina Bontcheva, and Peter Tolmie. Towards detecting rumours in social media. In *Workshops at the Twenty-Ninth AAAI conference on artificial intelligence*, 2015. 5.3.1

[143] Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, and Michal Lukasik. Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. *arXiv preprint arXiv:1609.09028*, 2016. 1.1, 5.2

[144] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie.

Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989, 2016. 5.2, 5.5