Hidden Fear: Evaluating the Effectiveness of Messages on Social Media

by

Mohsen Ahmadi

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2020 by the
Graduate Supervisory Committee:

Hasan Davulcu, Chair
Arunabha Sen
Baoxin Li

ARIZONA STATE UNIVERSITY

May 2020

ABSTRACT

The development of internet provided new means for people to communicate effectively and share their ideas. There has been a decline in the consumption of newspapers and traditional broadcasting media toward online social mediums in recent years. Social media has been introduced as a new way of increasing democratic discussions on political and social matters. Among social media, Twitter is widely used by politicians, government officials, communities, and parties to make announcements and reach their voice to their followers. This greatly increases the acceptance domain of the medium.

The usage of social media during social and political campaigns has been the subject of a lot of social science studies including the Occupy Wall Street movement (Conover *et al.* (2013)), The Arab Spring (González-Bailón *et al.* (2013)), the United States (US) election (Howard *et al.* (2017)), more recently The Brexit campaign. The wide spread usage of social media in this space and the active participation of people in the discussions on social media made this communication channel a suitable place for spreading propaganda to alter public opinion.

An interesting feature of twitter is the feasibility of which bots can be programmed to operate on this platform. Social media bots are automated agents engineered to emulate the activity of a human being by tweeting some specific content, replying to users, magnifying certain topics by retweeting them. Network on these bots is called botnet and describing the collaboration of connected computers with programs that communicates across multiple devices to perform some task.

In this thesis, I will study how bots can influence the opinion, finding which parameters are playing a role in shrinking or coalescing the communities, and finally logically proving the effectiveness of each of the hypotheses.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor Dr. Hasan Davulcu, who trusted me and supported me in all of my endeavours over our collaboration period.His endless energy and result oriented approach motivated me to deliver for every milestone of this journey. I am also thankful to my committee members; Dr. Baoxin Li, Arunabha Sen. The classes I have taken from them and their valuable feedback significantly helped me to shape this dissertation.

I can not be more grateful for my parents Fariba, and Alireza Ahmadi, and my siblings Mahsa and Poria. They were always a Skype call away to share the happiness of achievements and to uplift when things were not looking so promising. Without their endless passion for education and progress, I would not be where I am right now. Finally, I would like to thank my dearest colleague Ghazal Mousavi. Without her, without her endless support and thoughtful guidance , this journey would not succeed the way it is now.

Last but not least, I would like to give the most special thanks to Ava Shirzadeh. Without your endless love, support, and goofy jokes I would not be able to cope at this hard situation. Cheers to happier days ahead.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

PREFACE

This thesis has been written to fulfill the graduation requirements of the ASU School of Computing, Informatics, and Decision System Engineering (CIDSE). I was engaged in researching and writing this thesis from May 2019 to April 2020.

My background is coming from cybersecurity and lies in software security. I decided to work on this project because I noticed there is a surge in detecting propaganda and fake news in social media. Once social media has been considered as a medium providing freedom of speech, nowadays it is turned into a tool for spreading propaganda.

I could not have finished this project without a strong support from my family and friends. Thank you for unwavering support.

Chapter 1

COMMUNITY DETECTION AND VISUALIZATION

1.1 Introduction

Since its launch in October 2006, Twitter has become very popular among the people all around the world. Twitter is classified under microblogging mediums as a powerful mean for sharing what is happening now by exchange of quick, short messages. (Rosenstiel *et al.* (2015)) While social media once upon a time have been praised for increasing the awareness of people and providing the freedom of speech, nowadays became a source of spreading misinformation and propaganda. (Bessi and Ferrara (2016))

Social bots are growing rapidly on Twitter, but they are almost active in any other social media platform that is part of the political communications in countries. (Samuel (2015)) In this research we will study the Twitter as one of the wide spread base mediums used by most of the people and political parties around the world.

Back in 2006, Philip Howard expressed concerns about the possibility of public opinion manipulation, spread of political misinformation through social media. (Tewksbury (2007)) The degree of influence is highly dependent on how much the context of tweet is consistent with the priors of a human. For instance, a bot supporting the "leave" campaign has a stronger impact on a "leave" supporter than a "remain" supporter. Experiments (Bae and Lee (2012)) proved that the sentiment of tweets plays an important role in how the information is disseminated. For instance, a message with a positive sentiment generates another message with the same sentiment. In this work, we evaluate the effectiveness of a tweet based on considering different factors like

the positivity or negativity of the context, threats to the core values, and causality analysis.

The aforementioned result highlights the fact that people are tent to be part of communities of like minded people so that their beliefs are reinforced. This is exactly what is called the "echo chambers" (Colleoni *et al.* (2014); Quist *et al.* (2006)) effect in social media. Nowadays, political actors with adequate resources can deploy bots to attack opponents, amplify their values, spread their ideas. This means platforms like Twitter could enhance the separation and produces a more fragmented communities rather than a uniform one.

1.1.1 Community Detection

Social networks often described in the form of a complex network interconnected with nodes which resembles the entities or users in the community and edges are their connection. Identifying the communities in social media is a crucial part of social media analysis. (Girvan and Newman (2002)) A community is consist of a group of users that interact with each other more frequently and share similar interests. (Ozer *et al.* (2016)) Hence, the echo chamber effect of social media amplifies the polarization and segmentation in social media which create communities. Therefore, people are tend to be in groups of like minded people so their values are appreciated than being opposed.

Detecting communities in social networks attracted many researchers (Sánchez *et al.* (2016); Ozer *et al.* (2016)) to devise algorithms to find the best clustering. There are different metrics to assess the efficiency of a community detection algorithm like modularity.

Connections in the network or edges on the graph in Twitter can represents any of the follow, retweet, or user mention relationships. In our research, we don't rely on

follow information because in political domain it is unclear if it shows a support or opposition. (Myers *et al.* (2014)) study shows that follow graph on Twitter displays characteristics of both an information network and a social network. Intention behind following a news resource in Twitter has two folds, one can built on top of the social ties while on the hand could be for information consumption purpose only.

Among the retweet and user mentions, none of them has been proven to be a good indicator of approval. However, (Conover *et al.* (2011)) shows a practical case of applying mention and retweet interactions to find the the political polarization on Twitter. Therefore, we rely on populating the network using retweet relation, nodes as the users and the number of retweets happened between every two user as the designated edge weight.

Due to the size of the network, both in terms of nodes and links, I used Louvain clustering algorithm (Blondel *et al.* (2008)) to detect communities. The Louvain clustering is a method to extract communities from large networks using a greedy optimization on modularity. The computational complexity of the algorithm over a network of n nodes appears to be of order $O(n^2)$.

The idea behind the community detection algorithm is a greedy optimization strategy to optimize the quality function known as "modularity" of a network division. Based on the definition (Newman (2006)), modularity is the number of edges falling within communities minus the expected number of edges in an equivalent network if edges places randomly. This metric takes values inside the interval of $[-1, 1]$.

Louvain algorithm is divided into two phases that are repeated iteratively. First each node of the network is assigned to its own community which means in the initial run there are as many communities as there are nodes. Then, for each node i we consider its neighbours and we evaluate the gain of modularity by removing i from its own community and adding it to each of the neighbours' community. Then, node i is

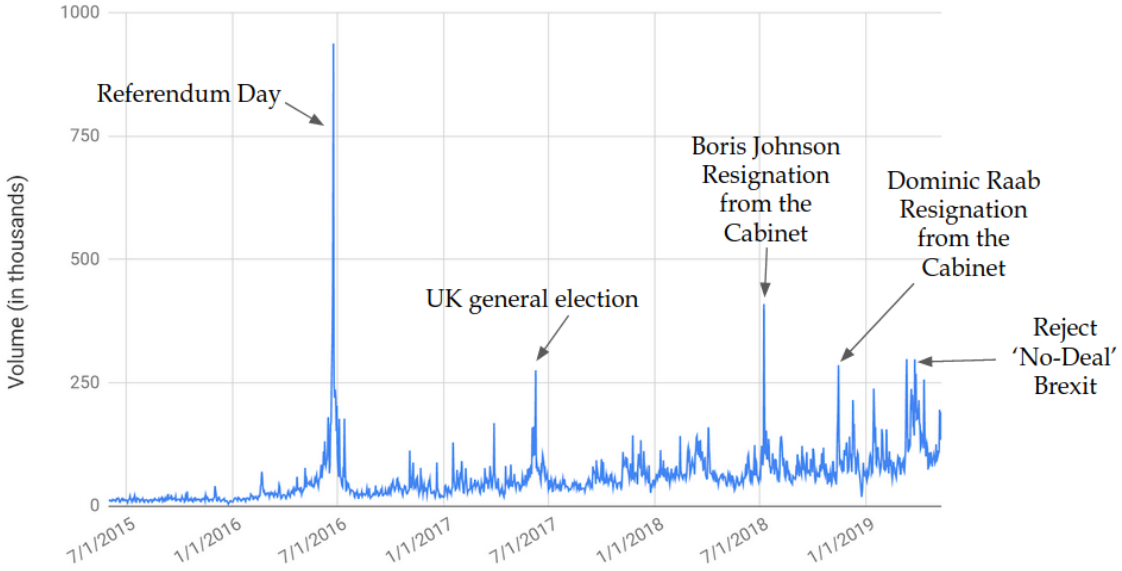


Figure 1.1: Distribution of Tweets in the Brexit Dataset

placed into the community for which the modularity gain is the maximum. The first phase will stop after reaching to the local optimum that happens when no individual node can be moved between communities to get a better result. The phase consists of building a new network by joining the communities from the previous step. The weight of the links between two new nodes will be computed by the sum of the weight of the links between nodes in the two corresponding communities (Blondel *et al.* (2008); Newman (2006)).

This thesis is focused on the Brexit and NS2 as two separate case studies. I will go through them in more details in Section 1.1.4. Given the huge amount of data we had, I decided to plot the distribution of tweets over the timeline. In Figure 1.1, we can see the trend of data and spikes in some specific periods for the Brexit. Using the information we gained from the Figure 1.1 we can find the list of breakouts. Breakouts are defined as date ranges at which we want to find the active communities on.

By running the Louvain clustering on the Brexit dataset, we found various communities that I break down some of them in Table 1.1. Among the breakouts the ones

Table 1.1: Results of Louvain Community Detection Algorithm on Brexit Dataset

Breakout	Date	Community	Cardinality	Clustering coef
		3	100999	0.0089
	2016-03-12	2	46172	0.1113
Referendum Day	-	7	82697	0.0338
	2016-08-21	0	65631	0.0548
		4	26183	0.0319
		4	44943	0.0810
	2017-04-27	5	87376	0.0083
UK General Election	-	2	55177	0.0282
	2017-11-06	0	59756	0.0511
		3	36874	0.1233
		4	94709	0.0361
	2018-06-08	2	94709	0.0361
B. Johnson Resignation	-	6	26986	0.00037
	2018-09-24	3	31424	0.1137
		0	28664	0.0874

that lies in one the the important events is of vital importance. In Table 1.1, I provided the results of running Louvain including the number of detected communities, cardinality (the number of nodes of edges in the network), and clustering coefficients. This table depicts communities of the Referendum Day (June 23, 2016), UK General Election (June, 8 2017), and Boris Johnson Resignation from the Cabinet (July 9, 2018).

1.1.2 Sankey Diagram

After finding communities, we need to visualize the network to get a sort of understanding of how these communities merge or shrink over time. To achieve this goal, we will use the Sankey diagram (Lupton and Allwood (2017)) to show the weighted network and their corresponding flows. Several nodes of the network represented by rectangles representing the communities. Their links are represented using arcs with different widths that are proportional to the flow.

A Sankey diagram visualizes the proportional flow between nodes within a network. The Sankey diagram we use in this study for the visualization exhibits similarities to "alluvial diagram" with marginal differences. The involved parameters that we match the flow against in our visualization is the cardinality or size of the community while an alluvial diagram visualizes the changes in the network over time. Although we depict the Sankey over a timeline, it's not involved the flow itself.

The networkD3 is the framework we used to create the Sankey that is based off the D3.js JavaScript library which allows users to create Sankey diagrams. To create the Sankey, we should provide parameters like Links and Nodes to this frameworks. Links is a data frame that includes the source and target of each link in the network. Nodes is a data frame containing nodes unique IDs and their corresponding properties.

The initial Sankey diagram of the Brexit detected communities and their respective flows over the breakouts is shown under Figure 1.2

1.1.3 Bot Detection

Bots have bee defined as automated agents that function on an online platform (Franklin and Graesser (1996)). Given the development of the internet and wide spread use of that, this term started to denote to a larger class of autonomous com-

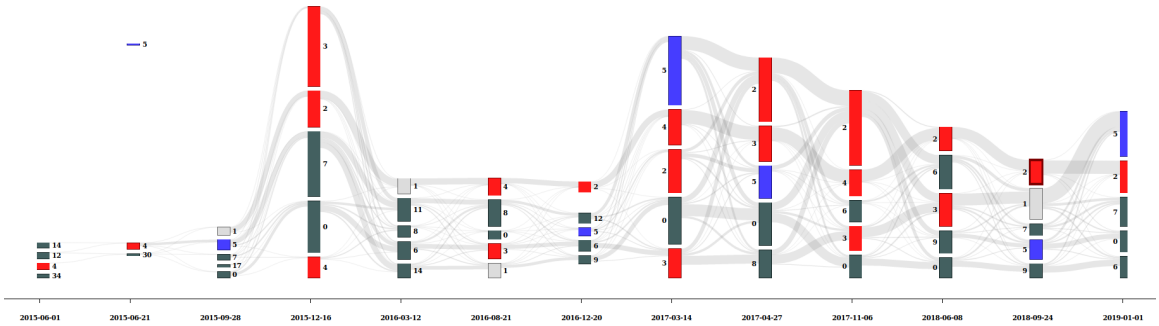


Figure 1.2: Initial Sankey Diagram Results Without Coloring the Parties

puter programs with different functionalities. One of the first studies of bots goes back to Leonard's paper when any automated script doing scraping, crawling, indexing, or even chat denoted as bot Leonard (1998).

Among various categories of bots in our study we focus on social bots. Social bot is a program that automatically generates contents and communicate with humans social media (Davis *et al.* (2016)). With the emergence of Twitter microblogging service in 2006 and the underlying Application Programming Interface (API), that encouraged the developers to create bots which are capable of engaging in discussions.

In 2010, researchers found that these bots could be used in balk for malicious purposes like spreading malware or spamming (Chu *et al.* (2010)). Social bots with specific purpose in spreading political content is called political bots (Howard and Woolley (2016)). One of the first use cases of political bots was during Massachusetts Special Election in the US by corrupting the reputation of one of the candidates (Metaxas and Mustafaraj (2012)).

Bots could increase the dissemination speed of a content by repeatedly bringing a matter to the attention of audience whether it is true or not. Online grassroots movements can be faked by masking the supporters of a message to make it appear as it originates from the movement leaders. This phenomena is called astroturfing and political actors and organizations deploy bots to spread misinformation and promote

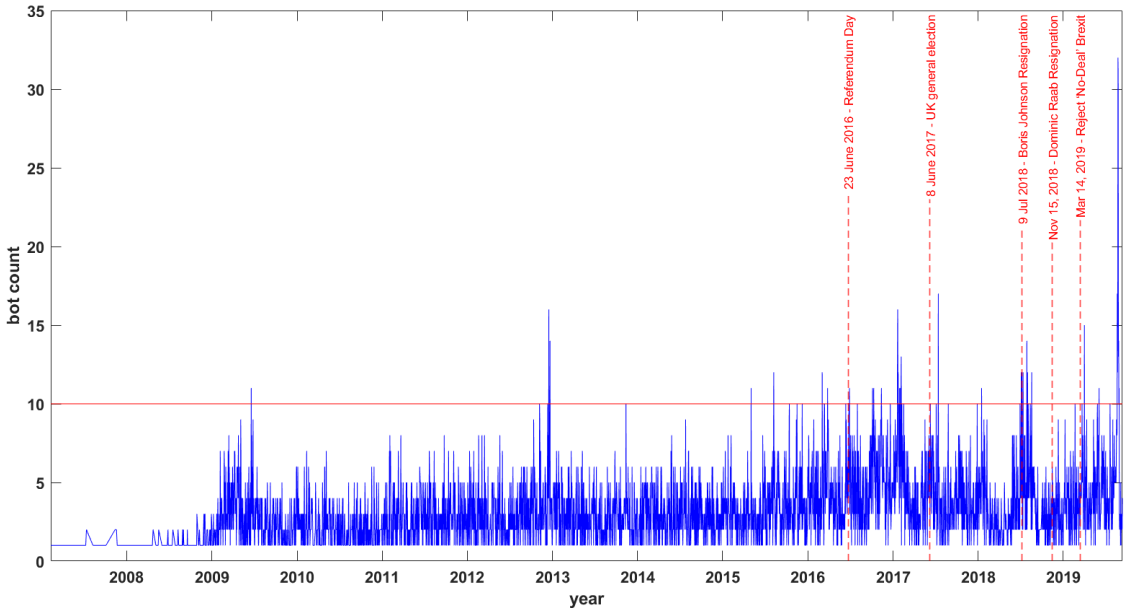


Figure 1.3: Bot Account Creation Dates/Rates

their own point of view (Llewellyn *et al.* (2019)).

While bots can be used by governments, political actors to spread their ideologies, it can also be used by covert agencies to spread misinformation and propaganda. There are some terminologies related to bots that needs to clarify. There is a term called "sock puppet" that is an account with a fake online identity used to interact with users on social media. The term draws from the manipulation of hand puppets by a human actor (Bastos and Mercea (2019)). Sock puppet refers to a general class of manipulative actors and includes bots, human (troll), or a combination of both (hybrid) (Bastos and Mercea (2019)).

One of the big actors of propaganda is the Internet Research Agency (IRA) Russian troll factory. The riffle false information dissemination (Jackson (2008)) context of IRA bots are inlined with the Kremlin foreign policy strategies with roots in Black Propaganda department of Soviet Union. The main goal of these agencies is to spread false information with an intention of deceiving public opinion and spreading the impression of chaos among people.

For the purpose of bot detection, we collected the initial list of bots from sources like the IRA-related trolls and bot accounts officially released by the Twitter. We use this list later on in Section 2.2 to color the bot infested communities in the Sankey diagram. In Figure 1.1, we can see the emersion rate of about 8000 bots over the years. It's obvious how quickly bots are evolving everyday and specially months before the important social or political venues.

Detecting the new bots is an ongoing labor intensive effort which starts with maps (listing known pro-Russian and other far/left or right fringe groups country-per-country), we detect fringe leaders and divisive/subversive issues and proceed to identify Russian bots that regularly (1) mention these issues and actors, (2) re-tweet at extra-ordinary high rates and (3) link to pro-Russian sources and their layering networks.

1.1.4 Dataset

We work on two different datasets in this study. The Brexit dataset contains nearly 51 million records of tweets in English about the campaign activities for events in the Pre-Brexit to some Post-Brexit events. There are two dominant political parties right-leanings and left-leaning.

Tweets were posted by 2.8 millions of unique users worth four years of data from June 1, 2015 to May 12, 2019. In detail, this dataset includes the records of tweets belonging to the Brexit for the duration of events that happened over the pre-Brexit, referendum day, the UK general election, Boris Johnson resignation from the cabinet, Dominic Raab resignation, and finally the post-Brexit deals.

Leave campaign is considered to be right-leaning and Remain is left-leaning. We could find 36 unique communities which 23 communities reside in the Remain and the rest are part of Leave party. Total number of unique tweets with Remain support-

ing content are 107, 700 with the total share of 38,255,498. On the other hand, there are 154,346 distinct tweets emphasizing on Leave values retweeted 41,535,64 times.

The second dataset is Nord Stream 2 (NS2) consists of around 8 million tweets collected using GNIP PowerTrack API[1] between November 6^{th} to January 26^{th}. The Nord Stream 2 is a new export gas pipeline running from Russia to Europe (Germany) across the Baltic Sea (Gazprom (2019)). In this debate there are two parties, one is anti-NS2 whom their actors are NGOs that criticise the NS2 on environmental, geological, and security issues. While proponents argues the pipeline is a key to Europe's supply security (Luke Sherman (2018)).

For the purpose of data collection, we have been querying the API with rules to search for hashtags covering both pro and anti parties. Selection of relevant hashtags chosen by a panel of academic experts. The list of hashtags and phrases was updated periodically, to reflect the evolving conversation on NS2. Selecting the tweets on the basis of hashtags has the benefit of capturing the contents that are most likely to be about the NS2 debate. This dataset has a combination of contents in various languages such like English, Russian, and Germany. In Table 1.2, we have a list of initial hashtags we used to scrub the relevant tweets for our analysis.

We used the IRA bots gathered from Brexit dataset on NS2. Based on phrases in the tweet content whether it is supporting anti or pro by checking the retweet network, we could classify tweeters and tweets accordingly. In detail, we had different phases to determine exactly which side a tweet is taking. (1) tweets contain specific pro/anti hashtag and phrases; (2) tweets that contain a link to a web resource like a news ar-

[1]GNIP is the first official Twitter data acquisition API owned by Twitter company in April 2014. It provides fast and easy access to the entire archive of Twitter data using PowerTrack filtering rules. It provides tweets and associated metadata including geo data, images, links, and mentions in a JSON format.

ticle, where the URL or the title of article contains keywords supporting or opposing sides; (3) retweets where it contains hashtag or phrases or the original text of the tweet; (4) retweets where the tweet content doesn't have the original tweet text, but it has a link refers to the original tweet.

In all there are 5,841 tweets opposing the NS2 consists of 4,412 users. On the other end, there are 516,050 tweets in support of NS2 campaign generated by 249,798 accounts.

Please note that we released all the datasets used in this research to the public. Email me at pwnslinger@asu.edu or it.ahmadi.91@gmail.com if you are interested in our dataset. We are currently sharing two labeled dataset of political Tweets for the Brexit and NS2 under the Twitter Agreement and Policies for content redistribution[2].

1.1.5 Data Cleaning

Twitter text is limited to a number of characters (140 characters in limited and 280 in extended version) which makes it people use abbreviations and slangs in their posts. The short content poses a problem in applying sentiment analysis on Twitter content. Additionally, emojis and URLs could introduce noise while analyzing the tweets. Therefore, there should be a set of passes that clean and screen tweets before we pass it forward for any further semantic mining.

First, for the tweets we exclude the special characters and tokens like links, and URLs from the text. Then, we remove mentions (starts with "@" sign) and retweet identifier following a mention (starts with "RT"). Second, since in most of the cases standard

[2]https://developer.twitter.com/en/developer-terms/agreement-and-policy - "If you provide Twitter Content to third parties, including downloadable datasets or via an API, you may only distribute Tweet IDs, Direct Message IDs, and/or User IDs. Academic researchers are permitted to distribute an unlimited number of Tweet IDs and/or User IDs if they are doing so on behalf of an academic institution and for the sole purpose of non-commercial research."

news media like BBC only post the link of news to their websites, we exclude tweets that only contained links or URLs in their content. Hashtags mostly used to support an idea or opinion and make the content noticeable by community however, in some cases where it is used in the middle of a sentence it carries meaning and is part of a speech. Therefore, as the last pass, we removed the hashtag sign ("#") and kept the rest as is.

Table 1.2: Pro/Anti NS2 hashtags and phrases used to harvest data using GNIP

Party	Hashtags	German Translations
Anti-NS2	#RethinkTheDeal, 4freerussia_org, Free Russia, #nordstream2, Opal pipeline, Gazprom, Russian energy, Russian LNG, expensive pipeline, corruption pipeline, money-laundering, US LNG, Shale gas, US freedom, freedom gas, energy security, methane leakage, captive of Russia, harmful to the environment, destroys the unity of the EU, Security threat to Europe, Security threat to EU, Russian propaganda, Russian weapon, Russian funded	#RethinkTheDeal, 4freerussia_org, Freies Russland, #nordstream2, Opal Pipeline, Gazprom, Russische Energie, Russisches LNG, Teure Pipeline, Korruptionspipeline, Geldwäsche, US LNG, Schiefergas, US Freiheit, Freiheit Gas, Energiesicherheit, Methanleckage, gefangen von Russland, umweltschädlich, zerstört Einheit EU, Sicherheitsbedrohung für Europa, Sicherheitsbedrohung für EU, Russische Propagandamedie, Russische Waffen, Russische finanziert
Pro-NS2	Nord Stream 2, Северный поток 2, NS2, natural gas, undersea pipeline, gas exports, US sanctions, Uniper BASF, Wintershall, gas as a weapon, energy choice, inexpensive gas, inexpensive energy, competitive gas, competitive energy, climate change, sustainable gas, reliable partner, stealing gas, Russian gas, energy transition	Nord Stream 2, Северный поток 2, NS2, Erdgas, Unterseeische Pipeline, Gasexporte, US Sanktionen, Uniper BASF, Wintershall, Gas als Waffe, Energie Wahl, Preiswertes Gas, Preiswerte Energie, Wettbewerbsfähiges Gas, Wettbewerbsfähige Energie, Klimawandel, Nachhaltiges Gas, Zuverlässiger Partner, Gas stehlen, Russisches Gas, energiewende

Chapter 2

COLORING THE SANKEY DIAGRAM

In the previous chapter, I covered the benefit of using Sankey diagram for the purpose of visualization of detected communities. In this chapter, I will cover a label propagation algorithm to spread the users' labels over the communities they belong to. Label propagation has been introduced as a fast reliable method in finding communities in a network of hundreds of thousands of nodes. We took advantage of algorithm discussed in (Raghavan *et al.* (2007); Xiaojin and Zoubin (2002)) to label the unlabeled data based on the already known list of labels in the network.

2.1 Label Propagation Algorithm

People want to discuss their opinions with the ones who are like-minded. One way to support an idea is to disseminate it over and over until it reaches to the hands of the proper audience. It is shown by (Ferrara *et al.* (2016)) that repeatability can give credibility to the content whether it is true or not. In political domain retweeting is a well-known way to show your support of a subject via sharing it with your network. This collaborative approach will keep the person in the loop and provides the opportunity for the person to receive similar messages from other members of the community in the future. Hence, labeling the retweet network can help us to answer the question of which party (label) a community is involved with.

In political domain, there are mostly two active parties called anti and pro. Which means we are usually dealing with two categories of labels, left-leaning and right-leaning. Existence of ideological segregation inevitably introduces a level of polarization in network. Therefore, it is easy to find high profile or influencer accounts

based on their follower network activity. One naive approach to find the influencers in Twitter is to sort the tweets based on descending order of retweet frequency. For example in the Brexit dataset, we manually labeled 311 top accounts with leave and remain labels.

The idea behind the algorithm (Raghavan *et al.* (2007)) is that each node in the community chooses to join the community to which the maximum number of its neighbours belong to. At each iteration, an unlabeled node gets its unique label and as the label propagation continues, densely connected nodes reach a consensus on a unique label. The label updating process is done asynchronously where the label of a node defined based on the label of its neighbours updated in this round and the neighbours has not been updated so far will take their respective labels from the previous iteration.

Our stop criteria happens when there is not node left changing its label. However, there could be nodes that have equal number of neighbours in two parties. In this situation, we break the ties among the two parties uniformly randomly. This way, there could be nodes changing their labels at every iteration while their neighbours labels remain constant. However, limiting the set of unlabeled nodes will reduce the range of possible solutions that the algorithm can produce. Seeding the algorithm with initial node labels improved the effectiveness and stability of execution.

The result of the label propagation algorithm will provide us labels for vertices in our detected communities. By applying the proper coloring for Sankey diagram, we can get the Figure 2.1 which distinguishes communities including the side they are taking. Furthermore, we enriched the Sankey with additional metadata like the top users, keywords, and bigrams. This could give us more insight into how communities deviate from Remain to Leave and vice versa. Also, which parameters or top keywords were important over those transmissions.

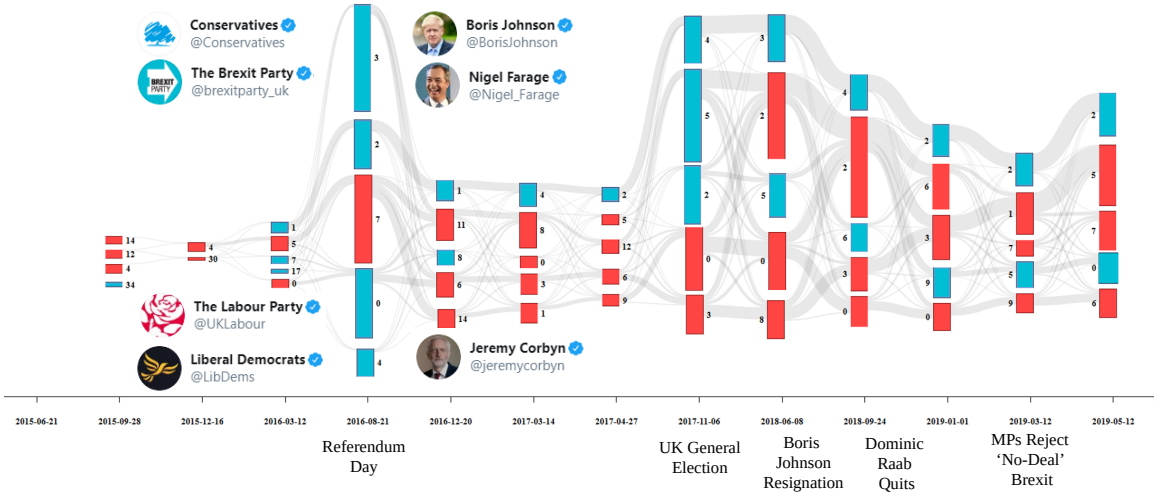


Figure 2.1: Enduring Leave/Remain Trends in Brexit, Blue Colored Communities Denotes Leave and Red Ones Belong to the Remain Campaign

The labeling algorithm pseudocode described in Algorithm 1. We defined the stop condition as an upper limit for the change ratio which implies the number of nodes changed their labels over all nodes. The other parameter is average threshold which signifies the label of an unlabeled node based on the average score of their neighbours. Since in this case we only have two labels, we set the labels as x and y in our logical notation that refers to leave and remain.

Label propagation is a greedy hill-climbing algorithm. It is highly efficient, but it can easily diverge to local optimum solutions depending on the initial labels assignment and random tie breaking (Conover *et al.* (2011)). By running the algorithm one hundred times, out of 1,430,833 users, 772649 are coded in Leave and 658184 are in Remain. The result is supporting the fact that UK votes to leave the EU (News (2020)).

Algorithm 1: Label Propagation Algorithm

Input : $G = (V, E)$ is the retweet graph, V_c is the set of labeled nodes
Output: $G' = (V', E)$ such that $|V| = |V'|$ and $V_c \subseteq V'$
begin
  $stop\_criteria \longleftarrow 0.01$
  $avg\_threshold \longleftarrow 0.5$
  for $v \in V$ **do**
    if $v \in V_c$ **then**
      $v.label \longleftarrow v_c.label$
      $v.score \longleftarrow v_c.score$
    else $v.label \longleftarrow None$
  end
  while $True$ **do**
    $next\_score \longleftarrow \emptyset$
    $change \longleftarrow 0$
    for $v \in V$ **do**
      if $v.label \neq None$ **then**
        $next\_score \longleftarrow next\_score \cup Pair(v, v.score)$
      else
        $next\_score \longleftarrow next\_score \cup Pair(v, calculate\_avg\_score(v))$
    end
    for $Pair(v, score) \in next\_score$ **do**
      if $score < avg\_threshold$ **then** $v.label \longleftarrow x$
      else if $score > avg\_threshold$ **then** $v.label \longleftarrow y$
      else $v.label \longleftarrow Random(x, y)$
      if $\neg v.score$ **or** $v.score \neq score$ **then** $change \longleftarrow change + 1$
      $V[v.id].score \longleftarrow score$
    end
    $change\_ratio \longleftarrow \frac{change}{|V|}$
    if $change\_ratio \leq stop\_criteria$ **then** break
  end
end

2.2 Coloring Bot Infestation

It is proven that social bots presence in political conversations has drastically evolved in a way they can mimic human behaviour which makes bot detection more difficult (Ferrara *et al.* (2016); Haustein *et al.* (2016)). Governors and political actors invest huge amount of money on creation of political bots. By taking a brief look at the

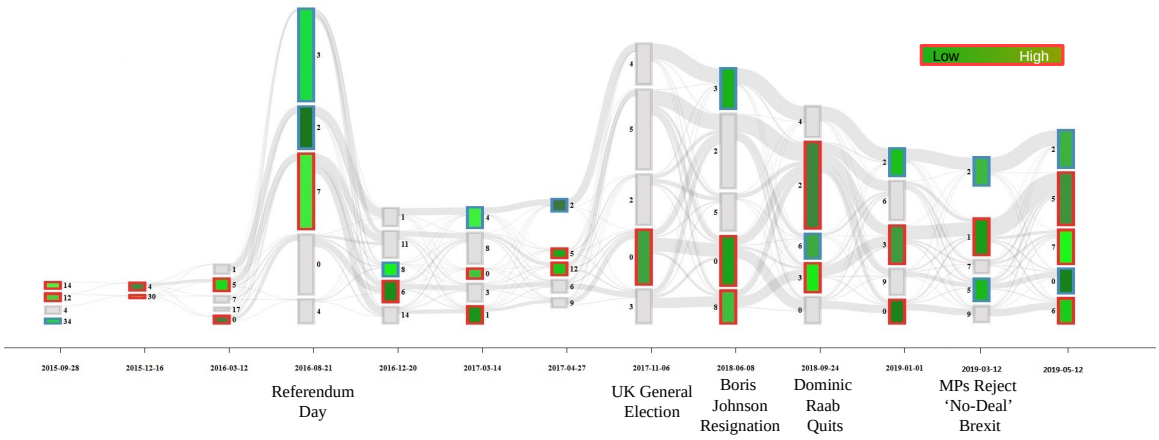


Figure 2.2: Bot Infested Communities in the Brexit Campaign

Figure 1.3, we can see the creation of bots just on or about the event announcement date and most of them are inactive before or after the event, but they massively generate content on event specific days.

In Figure 2.2, I depicted the infestation of bots in detected communities over the Sankey diagram. This can help us to study how and when bots pitching information to manipulate the public opinion and interfering in political discussions. It is quite common to see bots changing their direction by taking the side of another party to inject false news in the other party. In Figure 2.3, we can see the Sankey diagram of bot infested communities and the proportion each enduring community is involved with these bots/trolls in Pro-NS2 campaign.

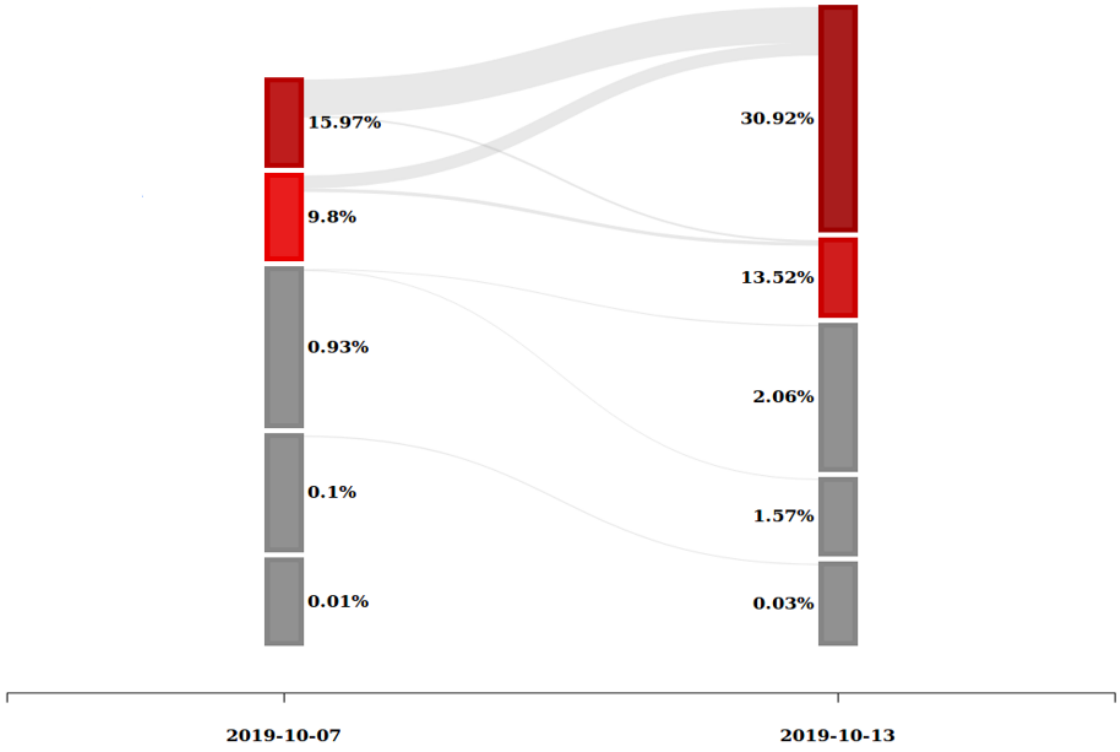


Figure 2.3: Bot Infested Communities in the Pro-NS2 Campaign. Labels by Each Community Shows the Percentage of Bots in That Community.

Chapter 3

MEASURING THE MESSAGE EFFECTIVENESS

In this chapter, I will cover three features helping us in understanding the effectiveness of messages. Per each introduced feature I will bring a hypothesis test to prove the validity of the hypothesis over the test population. We define effectiveness as a set of factors collectively affects the decision-making process of people. There has been research on finding the influences of negative framing on detecting political cynicism and politician accountability (Schenck-Hamlin *et al.* (2000)).

Frame connects news media messages to cognitive elements within the viewer. The framing effect happens when the frame interacts cognitive elements within the viewer, activating pa articular elements over others. The activated cognitive elements are going to influence the viewer judgements about that subject (Pan and Kosicki (1993)).

3.1 Sentiment Analysis

One important factor in dissemination speed of a tweet is the tonality or sentiment of the content. In political discourse, it is proven that messages with high emotional correlations spread faster among the users and has the potential to get trending. Based on (Kim and Yoo (2012)), the influence ratio of a tweet can be estimated using several indicators like (1) number of retweets (2) length of the discussion measured as the number of replies (3) number of people responded in retweet or reply chains (4) nesting degree of reply chains (5) durability of discussion.

Sentiment analysis or opinion mining is the task of automatically identifying the opinion of author about specific matter. One simple use case of sentiment analysis is to determine the polarity of words. Polarity determines whether a word is positive

or negatively framed. There are researches (Gorodnichenko *et al.* (2018)) relying on polarity and subjectivity score for the political content sentiment analysis. However, we found that Part-of-Speech (POS) tagging used in this research that is based on TextBlob (Loria *et al.* (2014)) is not efficient and comprehensive enough to support a variety of unknown Twitter POS.

To clarify it better, I will bring examples from (Gimpel *et al.* (2011)) on the Twitter-specific tags they introduced to support corner-cases happens in Twitter. For example, hashtags and at-mentions both can serve as a words or phrases in a tweet. Therefore, hashtags is tagged with their appropriate part of speech without considering the '#' sign. Mentions almost always are considered as proper nouns.

As another example, it is quite often to see apostrophes are omitted. As a result we encounter phrases like "ima" (slang version of I'm gonna) that could not be classified under any traditional POS categories. Tagger (Gimpel *et al.* (2011)) has introduced four new tags for words such like the previous example that are entangled together and cause confusion in detecting the right POS. To support the former example, they used Tag '**L**' that covers nominal + verbal form.

Sentiment usually is discussed in the body of polarity factor that shows how much positive or negative a phrase is. Also it can shows whether a phrase or word is neutral or not. Before I introduce the features, I would like to describe the hypothesis test and the Kolmogrov-Smirnov test I used to verify the null hypothesis.

3.2 Hypothesis Test

By definition, a hypothesis is an assumption about a population that may or may not hold. Hypothesis test is a formal method used to statistically prove whether to accept or reject a hypothesis. It is suggested to examine the entire population to see if a hypothesis is true or false. However, there are ways to sample the population

instead of using the entire set.

In order to test a hypothesis we need to find out whether it follows a discrete or continuous distribution. Understating the type of distribution is based on the variables. In our case, the distribution we based our study on is the number of times a set of tweets is retweeted. This distribution has a finite or countable number of values. It is possible to create a table that contains all possible values of retweeted tweets counts with their respective probabilities that sum up to one.

There are three different types of discrete distributions: Binary, Poisson, and Categorical. Since our variables are not of Binary type, we need to perform a goodness-of-fit test. There are two hypothesises should be defined for a test: Null hypothesis and Alternative hypothesis. Null hypothesis which is denoted by H_0, is a hypothesis that claims there is no significant difference between the sample population and the specified population. Alternative hypothesis which is denoted by H_1, is a hypothesis that claims there is a huge difference between the two populations.

In order to run a hypothesis test we need to perform the following steps: (1) stating the null and alternate hypothesis such a way that they are mutually exclusive. That means, if one is true the other should evaluates to false. (2) Creating an Empirical Cumulative Distribution Function (ECDF) for the two sample data. I will explain this in more detail at Section 3.2.1. (3) Passing the resulting two distributions to the chosen goodness-of-fit test. (4) analyzing the resulting p-value against significance level of test.

The results of a hypothesis test might introduce errors in the decision. There are two types of errors: Type I error occurs when we reject a null hypothesis when it was actually true. The probability of committing this error is called significance level and is often denoted by α. Type II error occurs when we fail to reject a null hypothesis that is false. The probability of *not* committing this error is called the Power of test.

The results of a test can be interpreted using two metrics: P-value and the region of acceptance. Suppose that S is the test statistic, then P-value is defined as the probability of observing the test statistic as extreme as S assuming the null hypothesis is true. The region of acceptance is a range of values such that if the test statistic falls within this range, the null hypothesis is accepted. The threshold for region is identified by the critical value which is calculated from a table given the significance level (α). The critical value is usually shown by $c(\alpha)$.

While the results of both approaches are equivalent, statistical textbooks use either P-value or region of acceptance. Using P-value, the null hypothesis rejects when the P-value is less than the significance level (α). The region of acceptance rejects the null hypothesis whenever the test statistic falls beyond the critical value ($c(\alpha)$).

In some notations we can see the resulting P-value preceded with one or trailing asterisks. Based on the American Psychological Association (APA) style, P-values less than or equal to the significance level ($\alpha = 0.05$) comes with one asterisk (*). Less than 0.01 are summarized with two asterisks (**) and below 0.001 are shown by three asterisks (***).

For the purpose of our experiments on hypothesis test, for each category of hypothesis, I generated a pair of distributions for Leave and Remain (Anti/Pro in NS2) campaigns for which the number of tweets are on the y-axis and number of shares are on the x-axis. Given the fact that we are dealing with a pair of discrete distributions on retweeted tweets count, we decided to use two-sample Kolmogrov-Smirnov (K-S test) goodness-of-fit test.

Since we didn't have a clear understanding on the underlying population distribution for our data, I decided to run K-S test. Although the K-S test has many advantages, the general implementation of this test cannot be used on discrete distributions.

Nevertheless, I implemented (Allen (1976)) which is a derivation of K-S test to support the discrete distributions as well.

The other benefit of using K-S test is that it is resistant to transforming the data points to logarithmic, reciprocals or any other transformation. A transformation will only stretch the frequency distribution, but will not change the maximum difference between two distribution.

3.2.1 Two-Sample Kolmogorov-Smirnov Test

The Two-sample Kolmogorov-Smirnov goodness-of-fit Test (Two-sample K-S test) is a non-parametric hypothesis test that evaluates the maximum absolute vertical difference between the Cumulative Distribution Function (CDF) of two distributions. If the difference is negligible, we can conclude that these two sample data are coming from the same distribution.

The Cumulative Distribution Function, $F(x)$ will give the fraction of sample observations that lies below x. If we sort all of the observations in ascending order, then the following holds:

$$\begin{aligned} y_1 \leq y_2 \leq ... \leq y_n \\ F(y_i) = \frac{i}{n} \end{aligned} \tag{3.1}$$

Suppose that the first sample is $X_1, X_2, ..., X_m$ of size m has the CDF distribution of $F_1(X)$ and the second sample is $Y_1, y_2, ..., Y_n$ of size n has the CDF distribution of $F_2(Y)$. We want to test the following hypothesis:

$$\begin{aligned} H_0 : F_1 = F_2 \\ H_1 : F_1 \neq F_2 \end{aligned} \tag{3.2}$$

The two-sample Kolmogorov-Smirnov statistic is:

$$D_{m,n} = \sqrt{\frac{m.n}{m+n}} \max_x |F_{1,m}(x) - F_{2,n}(x)| \tag{3.3}$$

In Section 3.2, I covered the relationship between test statistics and critical value and how P-value could help in whether rejecting or accepting a hypothesis.

3.2.2 Zero-Inflated Negative Binomial Regression Test

The distribution we based our study on is the number of tweets were retweeted exactly j times, where j is anything greater than or equal to zero. By analyzing the resulting dataset, I noticed an excessive amount of zeros on the tweet frequency column and two possible different processes that arrive at the zero outcome. Either the tweet content was not effective (based on the rules we will talk about it later on in Section 3.3) which is referred to as a "certain zero" outcome or it was effective which is a count process.

Figure 3.1, shows the histogram of the number of tweets were shared. By briefly looking at the histogram we can notice the data are over-dispersed and number of zeroes are inflated. Also, it is obvious that the variance of shared tweets are greater than the mean of shared tweets. This assures use that we can fit this distribution to a negative binomial model.

Since for some of the tweets the reason that the number of times no tweet were retweeted more than zero-time is the same reason as some other tweets were retweeted more than zero-time, the number of zeroes are inflated. Because of the prior, we can fit this dataset to a zero-inflated negative binomial regression model.

The zero-inflated model assumes that the zero number of tweets for a specific share count is due to two different processes. The zero-inflated negative binomial regression generates two models and combines them. The first model is a zero-inflated model that is a binary model usually a logit model to model which of the two processes is associated with a zero outcome. The second one is the negative binomial model that is the count model and predicts the count of tweets which are not certain zero.

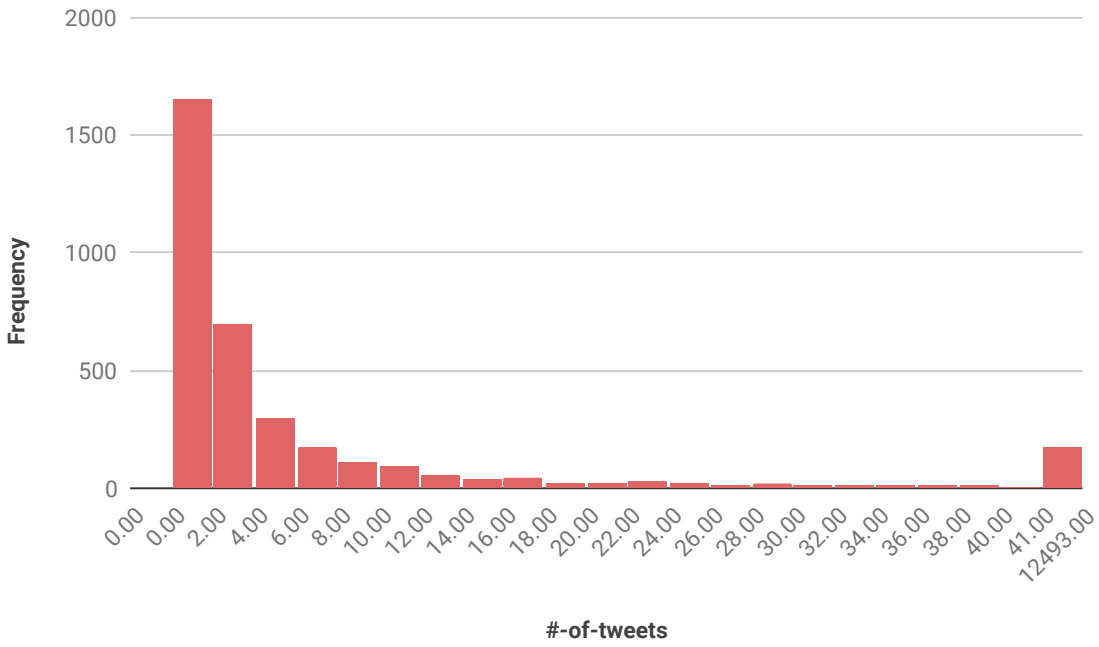


Figure 3.1: Frequency of Tweets Were Retweeted

Th reason behind designing a zero-inflated negative binomial regression test is to find which combinations of three hypotheses work together and results will show three of them were statistically significant in the model. In this model, the response variable is **shares_count** and I am exploring its relationship with **bot_count**, **follower_count**, $\mathbf{h_1}$, $\mathbf{h_2}$, and $\mathbf{h_3}$ variables. I am predicting count using **bot_count** and **follower_count** variables in the part of negative binomial model and predicting the certain zeroes using three variables $\mathbf{h_1}$ (negativity), $\mathbf{h_2}$ (causal inference), $\mathbf{h_3}$ (threats to the core values) in the logit part of the model.

3.3 Effectiveness Inference

In this section, I create a distribution based on the number of tweets were retweeted j times. I introduce three hypothesis on the distribution and I will evaluate each of them using a K-S goodness-of-fit test. Finally, I will run a zero-inflated negative bino-

mial regression model to see how the combination of these hypotheses work together.

3.3.1 Negativity

A positive response bring a positive reactions and a negative tone follows up with a negative feedback. Measuring the positive or negative influence of a phrase is called polarity. Polarity is a score between -1 to $+1$ which -1 is Negative and $+1$ is Positive. While (Gorodnichenko *et al.* (2018)) study shows that in the Brexit dataset messages with positive sentiment are the next-most prevalent after neutral. I this study, we show that for tweets bots/trolls engaged with the negative sentiments are marginally greater than the positive ones.

In our study we used the Tagger to find correct POS in our corpus. We collected the list of negative keywords by combining the moral-emotional negative keyword list (Brady *et al.* (2017)) with lexicons. Then we matched the tweets based on our new list of negative keywords using the polarity measure discussed in (Loria *et al.* (2014)). For the purpose of our experiments on testing the negativity hypothesis on the Brexit dataset, I generated a pair of distributions for Leave and Remain campaigns for which the number of tweets are on the y-axis and number of shares are on the x-axis.

Since our distribution is discrete and we are not dealing with categorical data, we cannot apply the chi-square test for the purpose of hypothesis test. The other option is to use a derivation of Kolmogrov-Smirnov goodness-of-fit test which is modified to support the discrete distributions. (Allen (1976)).

In political domain, retweeting a tweet is a way of giving credit to an idea and supporting that. By retweeting you are sharing the content of your interest with your network. In Figure 3.2 we can see the popularity distribution of negative and positive sentiments for Leave and Remain campaign in the Brexit dataset on a histogram chart. Also, in Figure 3.3, we have the histogram of popularity distribution of negative vs.

Table 3.1: Results of Running K-S Test for Hypothesis-1 on the Brexit and NS2

Dataset	Significance Level (α)	P-value	Test Statistic
Brexit Leave	0.05	9.5931e-05(****)	2547.87
Brexit Remain	0.05	5.9370e-79(****)	3582.04
Pro-NS2	0.05	1.85e-14(****)	139.74

non-negative sentiments in Pro-NS2.

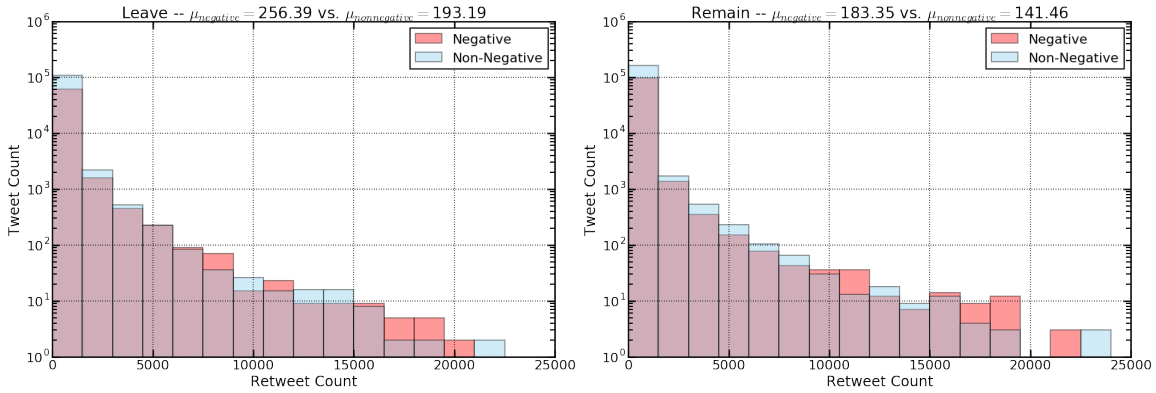
Equation 3.4 shows the definition of the null hypothesis (H_0) and the alternative hypothesis (H_1). Based on Table 3.1, since the P-value is far less than the significance level that we chose 0.05 ($\alpha = 0.05$), null hypothesis is rejected. From the results we can infer that there is a significant statistically difference between popularity distribution of tweets with negative sentiment and the popularity distribution of tweets with positive sentiment.

$$\begin{gathered} F_1 : \text{Popularity distribution of tweets with negative sentiment} \\ F_2 : \text{Popularity distribution of tweets with non-negative sentiment} \\ H_0 : F_1 = F_2 \\ H_1 : F_1 \neq F_2 \end{gathered} \tag{3.4}$$

Figure 3.4, shows the most negative framed phrases in the Brexit leave and remain.

3.3.2 Causal Arguments

A causal argument is a discussion in which parties bring reasons to support their argument by using causal keywords. To find a list of causal keywords, we manually annotated a list of keywords used in the discussions to negative, positive and neutral categories. In Table 3.2, I provided an expanded list of phrases we used to match against our dataset. We expanded the list using the synonym and acronyms from verb-sense of WordNet (Miller (1998)).



(a) Negative Framing in Brexit Leave (b) Negative Framing in Brexit Remain

Figure 3.2: Popularity Distribution of Tweets with Negative Sentiment Vs. Popularity Distribution of the Tweets with Non-negative Sentiment

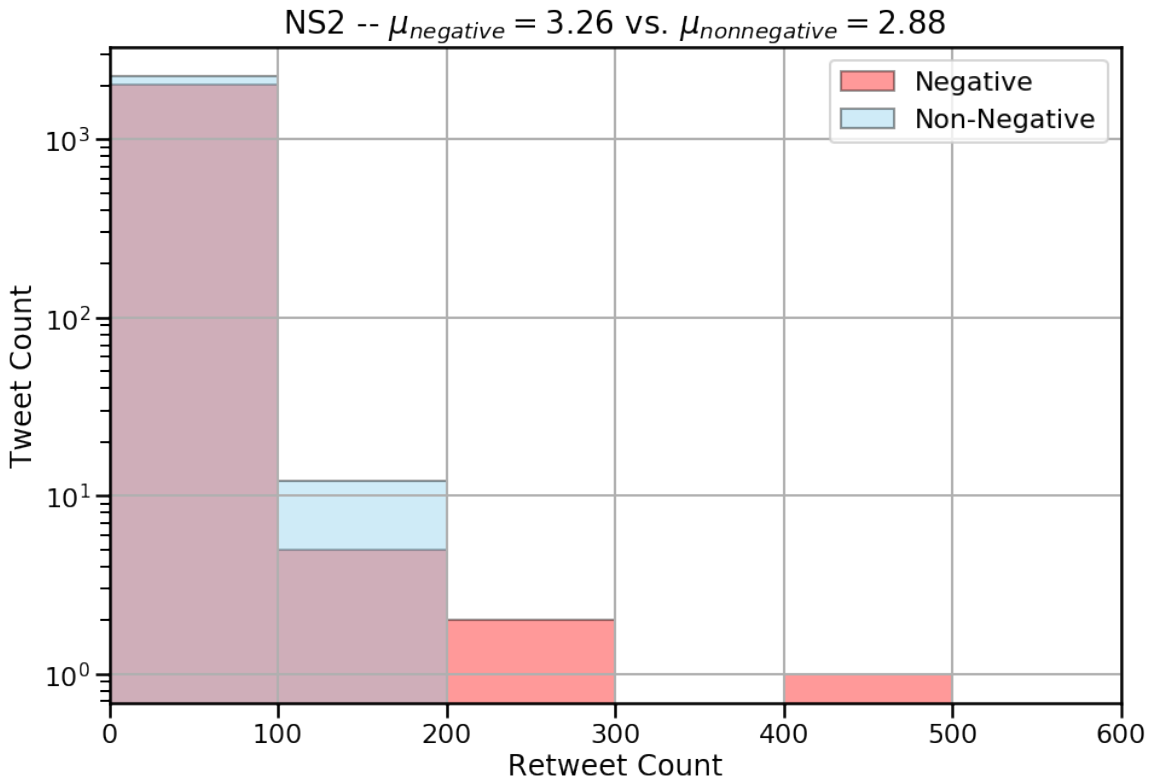
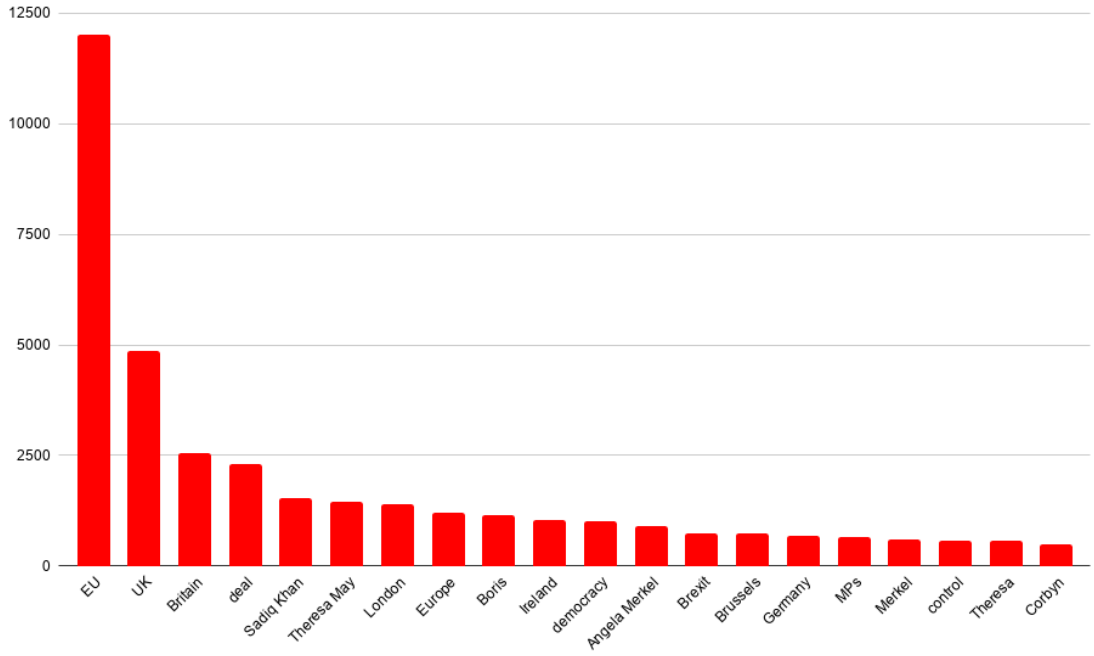
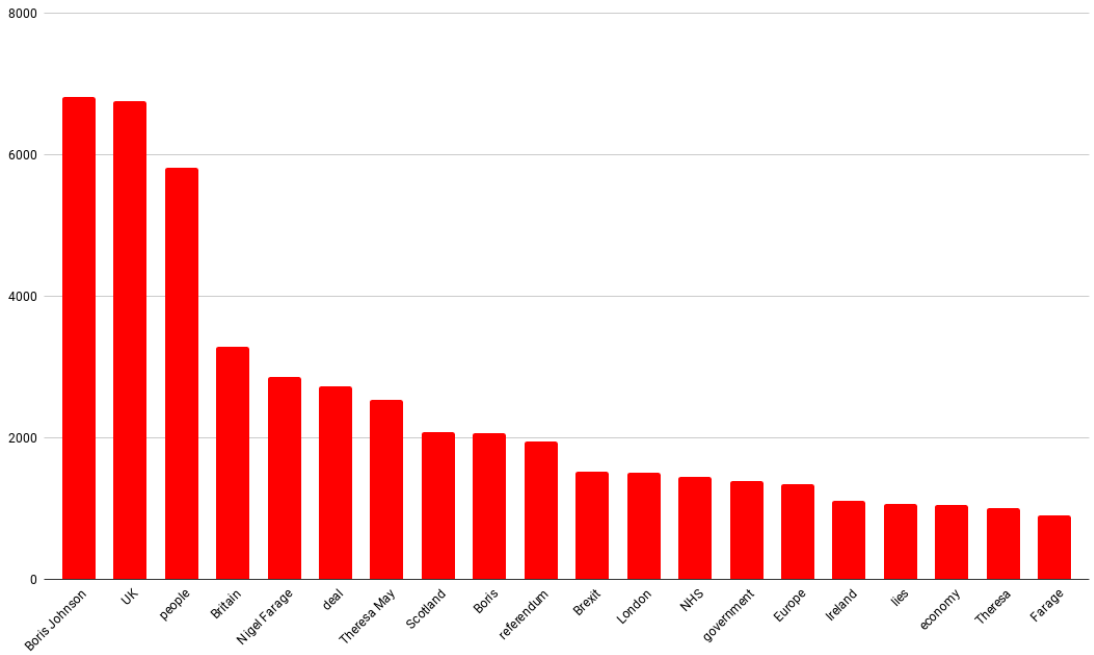


Figure 3.3: Popularity distribution of tweets with negative sentiment vs. non-negative sentiment in Pro-NS2



(a) Most Negatively Framed Entities in Leave



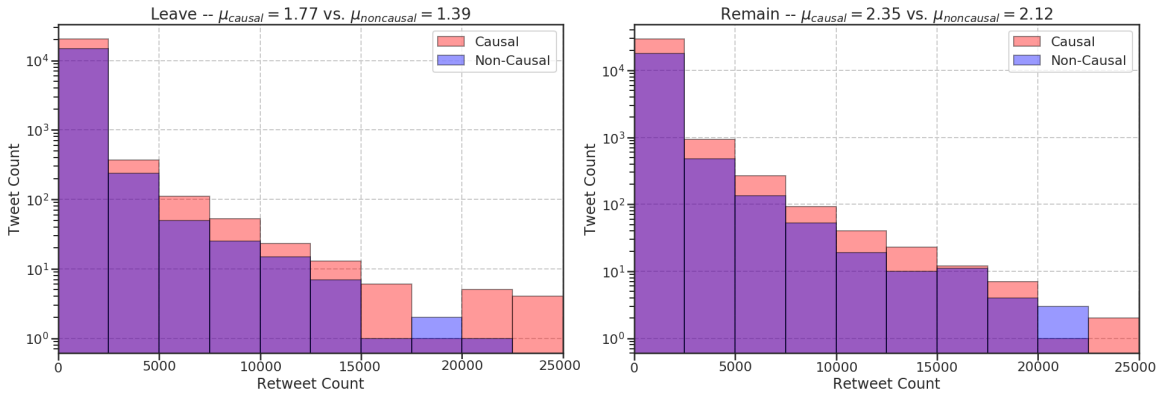
(b) Most Negatively Framed Entities in Remain

Figure 3.4: Most Negatively Framed Entities in the Brexit

Table 3.2: Expanded List of Synonyms and Antonyms for Causal Arguments

Polarity	Phrase List
Neutral	caused, causing, contributed, effect, expected to, impact on, implications for, implications on, led to, linked to
Positive	affect, affected, affected by, affecting, attributable to, result in resulted in, triggered, increased, promote
Negative	blamed for, decreased, destroyed, destroys, jeopardizing negatively impact, ravage, suffering from, thrown up by, halt exacerbated by, injure, killed, cripple

In Figure 3.5 we can see the popularity distribution histogram of the causal/non-causal arguments in Leave and Remain campaign. Also, in Figure 3.6, we have the histogram of popularity distribution of causal vs. non-causal debates in Pro-NS2. While we performed matching, any tweet with causal keywords from positive or negative category labeled as causal and the rest considered as non-causal. We ran the K-S test on the aforementioned distribution of retweeted tweets. To run the test, we defined the null hypothesis and alternative hypothesis as in Equation 3.5. In Table 3.3, since the P-value is far less than the significance level that we chose 0.05 ($\alpha = 0.05$), null hypothesis is rejected. From the results we can infer that the popularity distribution of tweets with causal arguments (F_1) do not follow the same popularity distribution of tweets with non-causal arguments (F_2).



(a) Causal Framing in Brexit Leave (b) Causal Framing in Brexit Remain

Figure 3.5: Popularity Distribution of Tweets With Causal Arguments vs. Popularity Distribution of Tweets With Non-causal Arguments

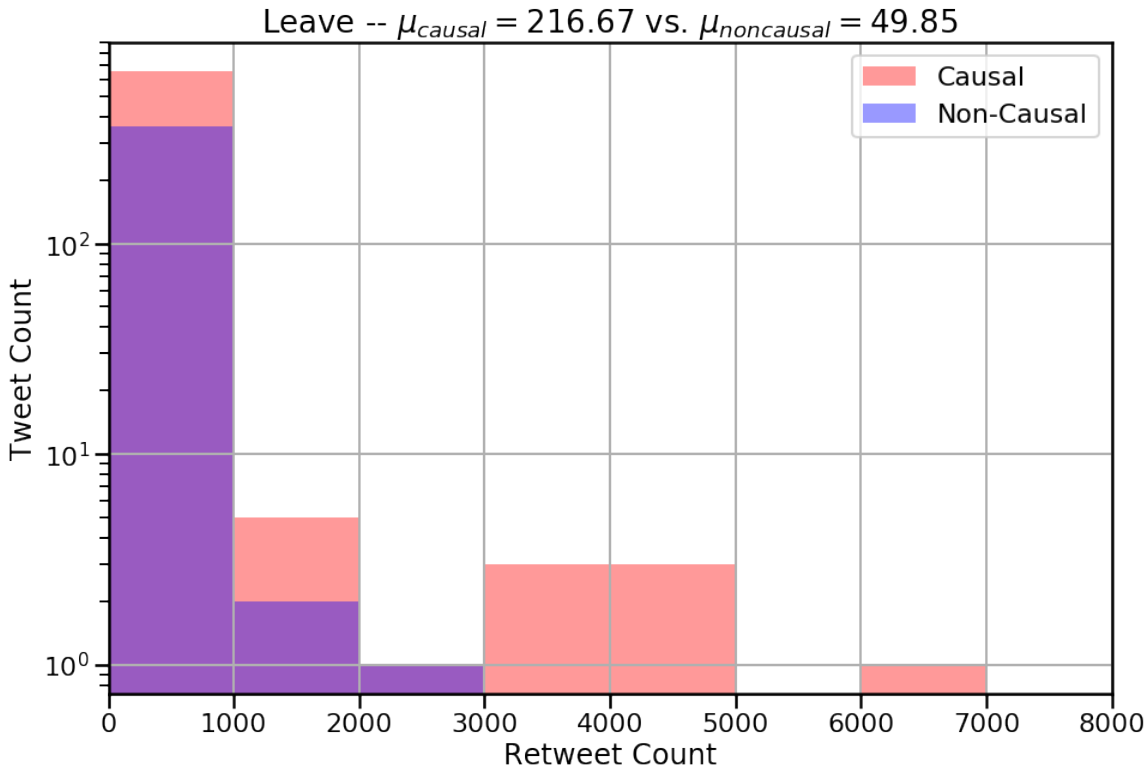


Figure 3.6: Popularity Distribution of Tweets With Causal vs. Non-causal Arguments in Pro-NS2

Table 3.3: Results of Running K-S Test for Hypothesis-2 on the Brexit and NS2

Dataset	Significance Level (α)	P-value	Test Statistic
Brexit Leave	0.05	5.41e-75(****)	494.85
Brexit Remain	0.05	7.19e-15(****)	179.58
Pro-NS2	0.05	0.00022(***)	99.58

$$
\begin{aligned}
&F_1 : \text{Popularity distribution of tweets with causal arguments} \\
&F_2 : \text{Popularity distribution of tweets with non-causal arguments} \\
&H_0 : F_1 = F_2 \\
&H_1 : F_1 \neq F_2
\end{aligned}
\tag{3.5}
$$

3.3.3 Threats to the Core Values

By precisely inspecting some threats of tweets, we noticed there are bots from left-leaning campaign actively posting and engaging into arguments in right-leaning campaign and vice versa. In order to attack a community, one effective strategy is to attack the weak link in the chain and vulnerable members. Since these members do not have enough knowledge they can easily get convinced and they can under question their believes by spreading propaganda and fake news. To make this happen, one need to find the values of an organization.

To get an understanding of values, we coded around 150 most frequently used keywords in negatively framed and negative causally liked tweets. Then we asked a group of experts in political domain to highlight the values with their respective polarity category. In Table 3.4, we can see a list of coded values with their respective category (Negative, Positive, Neutral). We also divided the values in Normative or Sacred sub-category. Values like country, religion, and nationality considered as sacred values and things like diversity, human, and healthcare are normative values.

After preparing the list of values, we defined a set of rules to find tweets with threatening content with respect to the values. A tweet contains a threat to a core value if it falls in any of the following category of rules:

- A tweet mentions a positive value and negatively framed
- A tweet mentions a positive Value and contains negative causal keywords
- A tweet mentions a negative Value and not-negatively framed
- A tweet mentions a negative value and contains positive causal keywords
- A tweet mentions a negative value and negatively framed
- A tweet mentions a neutral value and contains negative causal keywords

By performing a matching on tweet contents any tweet followed one of the above rules labeled as causal and the rest considered as non-causal. We ran the K-S test on the aforementioned distribution of retweeted tweets. To run the test, we defined the null hypothesis and alternative hypothesis as in Equation 3.6. In Table 3.5, since the P-value is far less than the significance level that we chose 0.05 ($\alpha = 0.05$), null hypothesis is rejected. From the results we can infer that the popularity distribution of tweets with threats to core values discussions (F_1) do not follow the same popularity distribution of tweets with non-threatening content. (F_2).

Figure 3.7 depicts the histogram of tweets with/without threats to the values for Leave and Remain campaign in the Brexit dataset. Also, in Figure 3.8, we have the histogram of popularity distribution of threatening vs. non-threatening discussions in Pro-NS2.

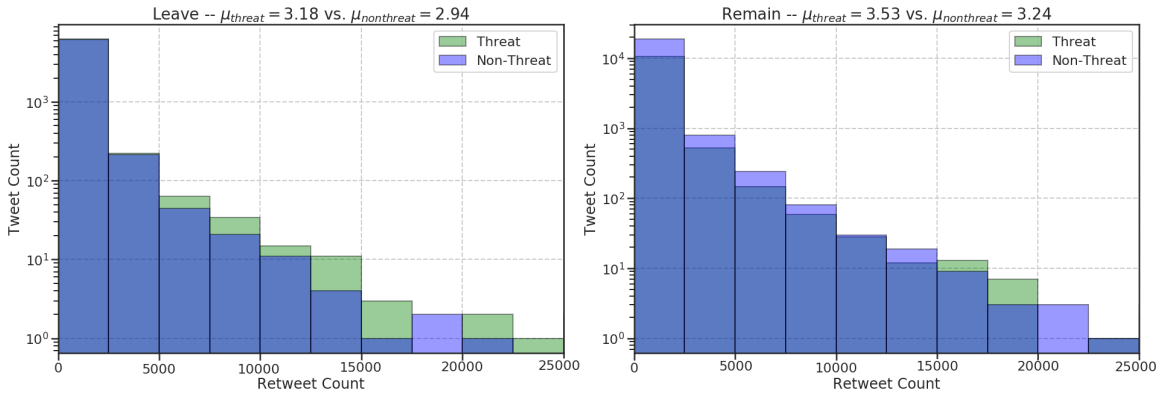
Table 3.4: List of Coded Values With Their Respective Polarity and Category

Polarity	Category	Phrase List
Neutral	Sacralize	African, America, American, Bishops, Britain, Canadians, Christ, Christian, Europe, God, Culture, Iranian, Jesus
	Normative	Brussels, Capitalism, Environment, Forests, Glaciers, Healthcare Human, Immigrants, Immigration, Job, Land, Water
Positive	Sacralize	Democracy, Independence, Liberty, Life, Security
	Normative	Cease-fire, Choice, Choices, Diversity, Equality, Health Honest, Justice, Nationalism, Nationalist, Peace, Patriotism
Negative	Sacralize	-
	Normative	Corruption, Deforestation, Hunger, Illegal, Insecurity, Lie Pollution, Sanctions, Wealth-inequality

Table 3.5: Results of Running K-S Test for Hypothesis-3 on the Brexit and NS2

Dataset	Significance Level (α)	P-value	Test Statistic
Brexit Leave	0.05	0.05(*)	59.24
Brexit Remain	0.05	4.58e-7(****)	116.73
Pro-NS2	0.05	3.88e-09 (****)	50.41

$$
\begin{aligned}
&F_1 : \text{Popularity distribution of tweets with threats to values discussions} \\
&F_2 : \text{Popularity distribution of tweets with non-threats to values discussions} \\
&H_0 : F_1 = F_2 \\
&H_1 : F_1 \neq F_2
\end{aligned}
\tag{3.6}
$$



(a) Threatening Framing in Brexit Leave (b) Threatening Framing in Brexit Remain

Figure 3.7: Popularity Distribution of Tweets with Threats to Values Discussions Vs. Popularity Distribution of the Tweets with Non-threats to Values Discussions

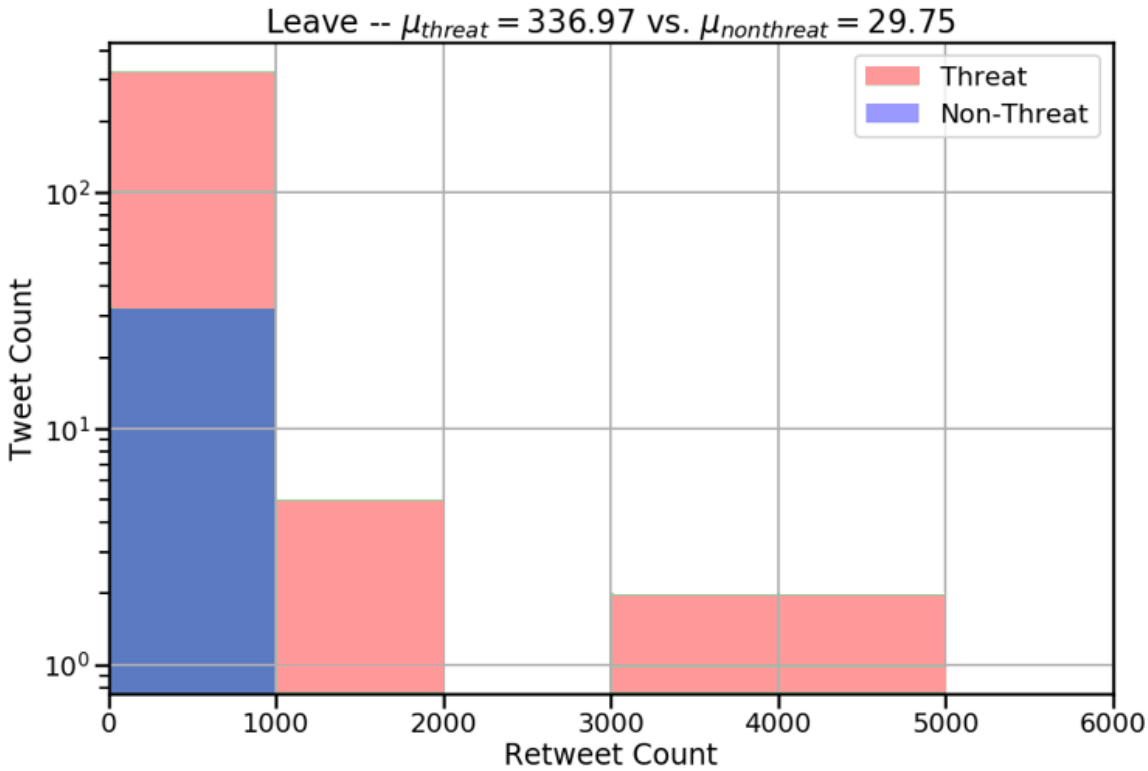


Figure 3.8: Popularity Distribution of Tweets with Threats Vs. Non-threats Discussions in Pro-NS2

3.3.4 Joint Effect

So far we have studied the outcome of negativity, causal argument, and threats to values in the effectiveness of tweets linked to the bots. The question left is the joint effect of these hypotheses on our distribution. To answer this question, I will use the zero-inflated negative binomial regression model discussed in Section 3.2.2.
In this experiment, my regression target variable is retweet count (*retweet_count*), and my predictor variables are as follows:

- *bot_count* quantifies how many bots are active in our dataset
- *follower_count* quantifies the number of the followers
- h_1 is the negatively framed tweets (hypothesis-1)
- h_2 is the tweets with causal argument content (hypothesis-2)
- h_3 is the tweets contains discussions on threats to a value (hypothesis-3)

From Figure 3.6 we can see that all of the predictors in the count and inflation portions of the model are statistically significant. This model fits the data significantly better than the null model (Intercept-only model).

Table 3.6: Results of Running the Zero-Inflated Binomial Regression

Campaign	Variable	Estimate	Lower 95%	Upper 95%
Brexit Remain	h_1	0.2977	0.2498	0.3455
	h_2	0.0427	0.0124	0.0731
	h_3	0.1388	0.1233	0.1544
	bot_count	0.8516	0.8405	0.8627
	follower_count	0.0	0.0	0.0
Brexit Leave	h_1	0.5588	0.4923	0.6252
	h_2	0.0436	0.0019	0.0853
	h_3	0.145	0.1265	0.1634
	bot_count	0.6765	0.6614	0.6917
	follower_count	0.0	1.6337e-5	2.1061e-5
Pro-NS2	h_1	0.5976	0.5301	0.6651
	h_2	0.5041	0.4713	0.5369
	h_3	0.1419	0.1176	0.1662
	bot_count	0.0052	0.004	0.0064
	follower_count	2.4626e-6	2.2728e-6	2.6524e-6

Chapter 4

CONCLUSION

4.1 Summary of Contributions

In this thesis, I proposed two new metrics to measure the effectiveness of messages on social media. I ran my experiments on the Brexit and Nord Stream 2 (NS2) tweets dataset. I provided mathematical prove for each of the hypotheses I mentioned. I used two-sample Kolmogorov-Smirnov goodness-of-fit test to prove the validity of hypothesis on the distribution of retweeted tweets. Finally, I used the zero-inflated negative binomial regression to see how the combination of these hypotheses works jointly. Given the results, I concluded that all of the predictors of model are statistically significant.

4.2 Future Directions

One of the limitations of this work is that I could not cover the sarcastic conversations in my rules. One research direction would be constructing rules or models that could capture the sarcasm in the social media and specially microblogging services like Twitter.

BIBLIOGRAPHY

Allen, M. E., "Kolmogorov-smirnov test for discrete distributions", Tech. rep., NAVAL POSTGRADUATE SCHOOL MONTEREY CA (1976).

Bae, Y. and H. Lee, "Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers", Journal of the American Society for Information Science and Technology **63**, 12, 2521–2535 (2012).

Bastos, M. T. and D. Mercea, "The brexit botnet and user-generated hyperpartisan news", Social Science Computer Review **37**, 1, 38–54 (2019).

Bessi, A. and E. Ferrara, "Social bots distort the 2016 us presidential election online discussion", First Monday **21**, 11-7 (2016).

Blondel, V. D., J.-L. Guillaume, R. Lambiotte and E. Lefebvre, "Fast unfolding of communities in large networks", Journal of statistical mechanics: theory and experiment **2008**, 10, P10008 (2008).

Brady, W. J., J. A. Wills, J. T. Jost, J. A. Tucker and J. 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).

Chu, Z., S. Gianvecchio, H. Wang and S. Jajodia, "Who is tweeting on twitter: human, bot, or cyborg?", in "Proceedings of the 26th annual computer security applications conference", pp. 21–30 (2010).

Colleoni, E., A. Rozza and A. Arvidsson, "Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data", Journal of communication **64**, 2, 317–332 (2014).

Conover, M. D., E. Ferrara, F. Menczer and A. Flammini, "The digital evolution of occupy wall street", PloS one **8**, 5, e64679 (2013).

Conover, M. D., J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer and A. Flammini, "Political polarization on twitter", in "Fifth international AAAI conference on weblogs and social media", (2011).

Davis, C. A., O. Varol, E. Ferrara, A. Flammini and F. Menczer, "Botornot: A system to evaluate social bots", in "Proceedings of the 25th international conference companion on world wide web", pp. 273–274 (2016).

Ferrara, E., O. Varol, C. Davis, F. Menczer and A. Flammini, "The rise of social bots", Communications of the ACM **59**, 7, 96–104 (2016).

Franklin, S. and A. Graesser, "Is it an agent, or just a program?: A taxonomy for autonomous agents", in "International Workshop on Agent Theories, Architectures, and Languages", pp. 21–35 (Springer, 1996).

Gazprom, "Nord stream 2: A new export gas pipeline running from russia to europe across the baltic sea", URL `https://www.gazprom.com/projects/nord-stream2/` (2019).

Gimpel, K., N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan and N. A. Smith, "Part-of-speech tagging for twitter: Annotation, features, and experiments", in "Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2", pp. 42–47 (Association for Computational Linguistics, 2011).

Girvan, M. and M. E. Newman, "Community structure in social and biological networks", Proceedings of the national academy of sciences **99**, 12, 7821–7826 (2002).

González-Bailón, S., J. Borge-Holthoefer and Y. Moreno, "Broadcasters and hidden influentials in online protest diffusion", American behavioral scientist **57**, 7, 943–965 (2013).

Gorodnichenko, Y., T. Pham and O. Talavera, "Social media, sentiment and public opinions: Evidence from# brexit and# uselection", Tech. rep., National Bureau of Economic Research (2018).

Haustein, S., T. D. Bowman, K. Holmberg, A. Tsou, C. R. Sugimoto and V. Larivière, "Tweets as impact indicators: Examining the implications of automated "bot" accounts on t witter", Journal of the Association for Information Science and Technology **67**, 1, 232–238 (2016).

Howard, P. N., G. Bolsover, B. Kollanyi, S. Bradshaw and L.-M. Neudert, "Junk news and bots during the us election: What were michigan voters sharing over twitter", CompProp, OII, Data Memo (2017).

Howard, P. N. and S. Woolley, "Political communication, computational propaganda, and autonomous agents-introduction", International Journal of Communication **10**, 2016 (2016).

Jackson, N., "'scattergun'or'rifle'approach to communication: Mps in the blogosphere", Information Polity **13**, 1-2, 57–69 (2008).

Kim, J. and J. Yoo, "Role of sentiment in message propagation: Reply vs. retweet behavior in political communication", in "2012 International Conference on Social Informatics", pp. 131–136 (IEEE, 2012).

Leonard, A., *Bots: The origin of new species* (Penguin Books Limited, 1998).

Llewellyn, C., L. Cram, R. L. Hill and A. Favero, "For whom the bell trolls: Shifting troll behaviour in the twitter brexit debate", JCMS: Journal of Common Market Studies **57**, 5, 1148–1164 (2019).

Loria, S., P. Keen, M. Honnibal, R. Yankovsky, D. Karesh, E. Dempsey *et al.*, "Textblob: simplified text processing", Secondary TextBlob: Simplified Text Processing **3** (2014).

Luke Sherman, J. W., "Gas pipeline nord stream 2 links germany to russia, but splits europe", URL `https://bit.ly/2Ka0y3T` (2018).

Lupton, R. C. and J. M. Allwood, "Hybrid sankey diagrams: Visual analysis of multidimensional data for understanding resource use", Resources, Conservation and Recycling **124**, 141–151 (2017).

Metaxas, P. T. and E. Mustafaraj, "Social media and the elections", Science **338**, 6106, 472–473 (2012).

Miller, G. A., *WordNet: An electronic lexical database* (MIT press, 1998).

Myers, S. A., A. Sharma, P. Gupta and J. Lin, "Information network or social network? the structure of the twitter follow graph", in "Proceedings of the 23rd International Conference on World Wide Web", pp. 493–498 (2014).

Newman, M. E., "Modularity and community structure in networks", Proceedings of the national academy of sciences **103**, 23, 8577–8582 (2006).

News, B., "Eu referendum results", URL `https://bbc.in/2KcLkLw` (2020).

Ozer, M., N. Kim and H. Davulcu, "Community detection in political twitter networks using nonnegative matrix factorization methods", in "2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)", pp. 81–88 (IEEE, 2016).

Pan, Z. and G. M. Kosicki, "Framing analysis: An approach to news discourse", Political communication **10**, 1, 55–75 (1993).

Quist, D., V. Smith and O. Computing, "Detecting the presence of virtual machines using the local data table", Offensive Computing (2006).

Raghavan, U. N., R. Albert and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks", Physical review E **76**, 3, 036106 (2007).

Rosenstiel, T., J. Sonderman, K. Loker, M. Ivancin and N. Kjarval, "Twitter and the news: How people use the social network to learn about the world", Online at www. americanpressinstitute. org (2015).

Samuel, A., "How Bots Took Over Twitter", `https://hbr.org/2015/06/how-bots-took-over-twitter` (2015).

Sánchez, D. L., J. Revuelta, F. De la Prieta, A. B. Gil-González and C. Dang, "Twitter user clustering based on their preferences and the louvain algorithm", in "International Conference on Practical Applications of Agents and Multi-Agent Systems", pp. 349–356 (Springer, 2016).

Schenck-Hamlin, W. J., D. E. Procter and D. J. Rumsey, "The influence of negative advertising frames on political cynicism and politician accountability", Human Communication Research **26**, 1, 53–74 (2000).

Tewksbury, D., "New media campaigns and the managed citizen, by phillip n. howard", Political Communication **24**, 4, 448–449, URL https://doi.org/10.1080/10584600701641532 (2007).

Xiaojin, Z. and G. Zoubin, "Learning from labeled and unlabeled data with label propagation", Tech. Rep., Technical Report CMU-CALD-02–107, Carnegie Mellon University (2002).