Truth or Troll: An Automated Framework for Identifying Authoritarian Regime Trolls in Twitter

by

Sindhuri Chandrupatla

(Under the Direction of Ismailcem Budak Arpinar)

Abstract

In the age of information, social media plays a vital role facilitating information flow. Unfortunately players with political agenda intervene and affect the information flow by spreading disinformation, that in turn affects our day to day life considerably. One recent such incident is the use of trolls - players who are paid to spread disinformation by agencies with intention to meddle with political events - that has affected the information flow significantly. This thesis focuses on a crucial problem of identifying the trolls, especially on twitter, from regular accounts by analyzing the user behavior. To the end, we also propose a pipeline to identify tweets that may potentially belong to a troll account by classifying how likely the tweet is a misinformation.

Index words: Troll account identification, Twitter data analysis, Latent Dirichlet Allocation, Text Classification, Word2vec

Truth or Troll: An Automated Framework for Identifying Authoritarian Regime Trolls in Twitter

by

Sindhuri Chandrupatla

B.Tech, CMR Institute of Technology, Hyderabad, 2011

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Master Of Science

Athens, Georgia

2018

Truth or Troll: An Automated Framework for Identifying Authoritarian Regime Trolls in Twitter

by

Sindhuri Chandrupatla

Approved:

Major Professor: Ismailcem B Arpinar

Committee: Khaled Rasheed
Krzysztof J. Kochut

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2018

Truth or Troll: An Automated Framework for Identifying Authoritarian Regime Trolls in Twitter

Sindhuri Chandrupatla

August 2018

Acknowledgments

I would like to express my sincere gratitude to my advisor, Dr. Ismailcem Budak Arpinar for the continuous support and motivation toward the completion of my Masters research.

I would like to thank my parents for their immense support and continuous encouragement throughout the process of my completion of the program. This would not have been possible without them. Thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

While advances of Internet technology and the evolution of social media have considerably altered our way of life for good, they, as well, come with a hefty cost, primarily due to spread of disinformation by a few players involved. In recent years, the spread of disinformation has been quite prevalent, and in fact to the extent that it affects the overall information ecosystem. Players involved, in most cases spread disinformation in order to accomplish a political agenda of sorts, sometimes directly or at most times indirectly via coordinated set of users called *trolls*. In this thesis, we specifically focus on developing a method to isolate such troll accounts from the rest in *twitter*.

We focus our work primarily on identifying the troll twitter accounts that can possibly be linked to agencies that have been accused of meddling with the 2016 US Presidential elections. It has been believed that social media platforms such as twitter has been used extensively for propagating disinformation by concerned

agencies with the intention to affect the elections. From a set of recently released reports, it has been believed that such agencies have hired people from within the country, solely for the purpose of spreading misinformation, in order to affect the outcome of the then ongoing election. Such hired participants have been actively posting tweets that aim to spread disinformation, misrepresent data and often make politically biased statements. In addition, they also involve in re-tweeting tweets and articles of such sorts. There have been several reports [1] even recently, more than an year after the elections, that such troll accounts may still be actively participating in flooding the twitter with misinformation. While such troll accounts and the associated agencies have been reported to be associated with various other cases, we restrict the scope of our work to focus only on their possible interference and effects during the elections.

Over an year after the case of interest of our study, there has been several advancements concerning the issue. In 2016 twitter has removed over 200,000 troll tweets [2] that has been believed to be indulging malicious activities and are connected to the agencies that are accused of interfering with the elections. The report [2] also states that twitter has emailed 700,000 as they fear they may have exposed them to the trolls, and expect the numbers to be much higher. Subsequently, in February 2017 twitter have handed over a list of 3,814 accounts to the congress that are identified as trolls. While we are able to obtain the account names that are identified as trolls, unfortunately the tweets posted by the account are not available as twitter has deleted those accounts and also have removed the data associated with them. But [2] have been able to recover partial information

from those suspended accounts, and have made them available publicly. We use the available data to perform our analysis where we train classifiers to learn to identify troll accounts from true ones.

Despite identifying and removing several of such troll accounts by twitter, it is quite obvious that several more troll accounts continue to exist and spread disinformation regarding several other similar affairs. Identifying them is of at most importance given the damage that they have incurred over the period of time. The objective of this thesis: given a set of tweets posted by an account collected for over a few months, we analyze the behavior of the user, and provide a likelihood of (1) the user being a troll or not, and (2) how likely a given tweet is a misinformation.

Chapter 2

Related Work

In recent times, as social media has revolutionized the spread of information, in parallel, the spread of misinformation has surpassed the tolerable limits. Curbing Internet trolls or spammers is the need of the hour as these corrupt players have been affecting our day to day life, such as the Russian Trolls in the 2016 US Presidential Election. Fortunately, there has been a considerable progress in that direction. Fourney et al. [2017] proposed a method that analyses the traffic to websites that were explicitly known to spread fake news before the 2016 US election and concluded that social media has in fact played a vital role in spreading *fake news*, that in turn altered voting patterns considerably. It has also determined that the voting patterns was in correlation with the traffic to these known websites. Shao et al. [2017] studied the role of bots in spreading misinformation on 14 million tweets were collected both during the 2016 elections and after that. It made a key observation that bots in fact played a key role in spreading misinformation and

biased and untrue claims, which to be noted, that later on went viral due to re-tweeting in the targeted geographic locations. Similarly Allcott and Gentzkow [2017] explored the behavior of the sites known to spread fake news and how crucial a role it played in the elections for the sake of determining the effect of the fake news spread in the outcome of the election. Later on Howard et al. [2018] analyzed whether the spread of misinformation and its correlation with the swing states. They in fact concluded that there was a positive correlation between the fake news and the outcome of the elections.

Much after the election, NBC (National Broadcasting Company) [2] network published an article releasing the tweets and user information of the known Russian trolls released by Twitter. Not only the elections, but the trolls have been reported to continue to operate much after the elections. This is an article [1] published by NBC where it specifically discusses the role of Russian trolls in parkland shooting. Analysis of Zannettou et al. [2018] on Russian trolls during the elections, on several parameters such as geolocation, account evolution, content analysis such as top hashtags, top urls used. Though they concluded that the influence was limited, they also mention that the analysis was done only on partial data, and was done much earlier than when twitter released more accounts. Another such approach is Badawy et al. [2018] which in detail explored the role of trolls, that analyzed parameters like geolocation, discussion topics like red *vs.* blue, whether the account was a bot or not.

The recent availability of data from trolls accounts made the exploration easier. Galan-Garcia et al. [2016] did a supervised learning of collected tweets of cyber-

bullies and non-cyberbullies to identify fake profiles. Mihaylov et al. [2015] also focused on the very same problem of identifying trolls vs non-trolls by learning discriminative classifier using supervised data, but worked on a different case of interest in Bulgaria. Subsequently Mihaylov and Nakov [2016] explored a Logistic Regression-based method for profile detection.

Chapter 3

Datasets

3.1 Russian Trolls

In the wake of the developments in investigations on Russian meddling of the Presidential Elections 2016, especially pertaining the misuse of twitter as a platform to propagate misinformation, Twitter released an official list of 2752 accounts that were believed to be involved in the false propaganda, while the tweets posted by the account were not released by Twitter. The list of identified troll accounts, was then handed over to the Congress. Aftermath of the news, twitter also subsequently took action by suspending the user accounts and removing their tweets. While only the user profiles are available and the tweets by them are not, NBC was able to acquire a partial list of the tweets and have released them to public. To do that NBC hired three sources that are familiar with the twitter's data systems which were able procure the data from the suspended accounts. NBC then cross-

referenced the tweets and user accounts and to recover a partial list of 202,973 tweets from 454 twitter accounts from the list of the 2752 accounts released earlier by twitter, excluding any accounts that were mistakenly confused for a Troll profile and were later restored. The sources used by NBC, however, are to remain anonymous in order to avoid being identified to violating the twitter's developer policy. According to Dr. Jonathan Albright, Research Director Tow Center for digital Journalism, Columbia University, the procured database comprising the identified suspended accounts and their tweets, is one of the largest repositories of the deleted Russian Twitter troll activity till date.

Our goal is to learn the behavior of troll twitter accounts specifically pertaining to the presidential election 2016, we use the datasets collected by [2], which consist of 454 twitter accounts with 203,452 tweets, averaging 450 tweets for each twitter account —the tweets recovered by [2] after twitter has suspended the accounts and removed all associated data of the troll users. Among the total number of tweets 175,185 were unique, rest are repetitions or retweets. The tweets were collected between 2014-07-14 and 2017-9-26. Using the data, we learn to model the nature of tweets of the troll accounts, as well as the troll account behavior analysis. Sample tweets from the identified Russian Troll dataset is shown in figure 3.1.

3.2 Elections 2016

On the other hand, in order to model the nature of a regular tweet pertaining to the election 2016 —that does not belong to the troll user account —we use

the dataset collected and made available from Paragon Science. The tweets are collected during the 2016 elections, and are made publicly available. The dataset comprises of 1126 tweets from various users (after removing duplicates and troll user tweets). Sample tweets collected that are related to the US presidential elections of 2016 is shown in figure 3.2.

3.3 Elections 2018

In addition, we collected tweets for more recent activities. One of such events is the ongoing Midterm Elections 2018. Since there are speculations concerning the Russians could again meddle with the ongoing elections, we thought it will be an interesting exploration to gather twitter data of the ongoing process and evaluate the method. For that purpose, we have collected tweets from 2018-02-14 to 2018-07-05 using hashtags *midtermelections*, *bluewave* and *redwave*. These are the hashtags that are commonly trending in the twitter trends over the past few months, and are pertaining to the elections. In total we have collected 29,660 tweets from 11,930 unique users. Figure 3.4 shows sample of collected tweets that are relevant to the ongoing mid-term elections.

3.4 Parkland shooting

In addition, one of the recent incidents that had speculations that the Russian Trolls again tried to meddle with, this time, the public opinions is the parkland shootings. Parkland shootings is one of the most important incidents where there

Tweet No	Sample Tweets from Russian Troll Dataset
1	RT @mc_derpin: #TheOlderWeGet the more pessimistic we are https://t.co/zS3jHZJl8P
2	RT @dmataconis: Ready To Feel Like A Failure? Joan Of Arc Was Only 19 When She Was Burned At The Stake http://t.co/S2j1IFm4y9
3	Amen! #blacklivesmatter https://t.co/wGffaOqgzl
4	RT @NahBabyNah: Twitchy: Chuck Todd caught out there shilling for Hillary Clinton The post BUSTED: Adam Baldwi... https://t.co/ay28pMpDw6 #...
5	RT @mcicero10: #BernieSanders #Trump people should rally TOGETHER against the establishment who is 💩 -ing on both choices #thefix
6	RT @ItsJustJaynie: @HillaryClinton The undecided voters on that stage was polled and said trump won. @cnn is biased.
7	@TodayCleveland 'no way'
8	@NickTomaWBRE Hi, Nick! We're holding a "Miners for Trump" rally tomorrow. If you're interested in covering it, ple... https://t.co/mUGySkE4UR
9	What. Is. A. Resolution #My4WordNewYearsResolution
10	Lifetime movie your pet psycho neighbor = you digging a grave for Rover. #TVLifeLessonsILearned
11	RT @Conservatexian: New post: "UN alarm that most of al-Shabab's force in Somalia are kids" https://t.co/OzRhmln0f4
12	RT @HillaryClinton: This one's for you, Hillary. https://t.co/KtzpIpSziO
13	RT @leonpui_: Hillary Clinton, Obama and the Democrats use the Communist born word, "RACIST" to divert the real Problem! https://t.co/pSEL...
14	RT @neus2100: Obama's Legacy! https://t.co/8gnO56BKKk
15	5th grade. When the second plane hit, I thought our school was next #My911Story
16	#teapartynews #teaparty #theteaparty #politics #conservative https://t.co/OTeVnYHv07
17	RT @DMashak: #Debates Chris Wallace @FoxNewsSunday: Pls ask if the candidates subscribe to the agenda & tactics of Alinsky's #RulesForRadic...
18	RT @mikefdupjourney: @she_nutt You're welcome! :) https://t.co/L7GqJ3xZzt
19	RT @PrisonPlanet: Hillary's anti-Trump poster child Alicia Machado LIED about Trump causing her eating disorder. https://t.co/jU1ryskgXM
20	Sheriff Joe Arpaio to speak at RNC after all #politics

Figure 3.1: Example tweets from troll twitter accounts.

Tweet No	Sample Tweets from Elections 2016 Dataset
1	""What Are You Hiding?""",The world is witnessing the unthinkable reemergence of a mini Cold War between nuclear powers: increasingly ominous actions by the United States/NATO and Russia; dangerous tensions between
2	'She Was Using for Year' : People.com,"Hillary Clinton 's email scandal continues. On Friday the New York Times reported that Clinton told FBI officials former Secretary of State Colin Powell had advised her to use a personal email
3	2015 when she answered questions from reporters in Fort Dodge
4	2016 6:19pm TAKE 2. Reposting FYEE. Mr. Jorgensen Nonsense. If the one with the most money always wins
5	2016 @ 12:00 pm I cant believe all the pandering to a group of violent hateful inbred mental savages to the detriment of everyone elses human rights. Send them and their enablers packing Reply ronyvo August 25 2016 @
6	2016 @ 7:21am in #trump #clinton Heads are exploding at CNN as pundits try to define Trump's repeated offers to help African-Americans as typical Hitler behavior. As my regular readers know what we have here is a perfect trap
7	2016 Becky Rosati Better wake up AMERICA.....do you want a ruler telling what we are going to be allowed to do? Please vote Trump Reply August 24 2016 Robert Tipsword Very well said Becky Reply August 24
8	2016 Issue - Trump vs Establishment and 3rd Party Candidates The world is witnessing the unthinkable reemergence of a mini Cold War between nuclear powers: increasingly ominous actions by the United States/NATO
9	2016 at 7:02 pm This is my California! The CA the Media refuses to show & lies about. Do you see this?!?🇺🇸 🇺🇸 #MAGA #LatinosForTrump pic.twitter.com/ssE1biLAga — FluffyDogAttack (@FluffyDogAttack) June 7 2016 Like
10	2016 at 8:38 am Did anyone who watched it hear the lid pop? If it didnt pop then it was a staged jar that was already opened... Reply August 24
11	2016 referenced unfounded rumors and innuendo regarding Melania Trump wife of Republican Presidential candidate Donald J. Trump
12	@CNN seems to be suffering a ratings collapse. What should we do? > #BlackOutCNN #Trump https://t.co/hVANFeZd5b""",When you tweet with a location, Twitter stores that location. You can switch location
13	Assange accused Clinton of engendering a "kind of neo-McCarthyist hysteria," referring to Joe McCarthy,,,
14	08/22/18, 2016 - YouTube,"""The American Dreamer"" Make America Great Again! Make America Safe Again! - Duration: 4:28. 282 views 4:28 Nigel Farage on Fox after speaking at Donald Trump Rally in Mississippi - Duration:
15	Blames Roger Ailes - ABC News Ann Coulter did not hold back when talking about Roger Ailes on this week's episode of " Powerhouse Politics " podcast. Coulter who this week published her book
16	But Still Heavily Favored in Utah - Public Policy Polling,In a year where much has been made of voters not liking their choices for President, there's probably nowhere that's truer than Utah. Donald Trump has a 31/61 favorability rating
17	CBS," and NBC`s morning and evening newscasts on Wednesday all punted on mentioning USA Today`s above-the-fold scoop about how an """"Istanbul-based college professor...accused by the Turkish government of coordinating
18	Chernin Group founder Peter Chernin MarketShare CEO Jon Vein
19	Clinton's State Department significantly increased arms export authorizations to the country's autocratic government even as that nation moved to crush pro-democracy protests . In a statement quoted by the Wall Street
20	Democratic Party Chair Francine Busby touted the merits of Trump-shaming. "Republicans in San Diego County are either making excuses for Trump scrambling for cover

Figure 3.2: Example tweets related to the presidential election 2016.

Tweet No	Sample Tweets from Parkland Shooting 2018 Dataset
1	@Emma4Change I like the cause but I wish they would march through the communities in our city that have the problem and not block people commuting.
2	@davidhogg111 @Emma4Change #EnoughIsEnough https://twitter.com/EdKrassen/status/1017205116565323776 …
3	@Emma4Change @openingtirade @WayneCh86085175 I understand.<Emoji: Victory hand (medium skin tone)>
4	@Emma4Change @WayneCh86085175 @openingtirade What does your comment has what to do with mass shootings NOT perpetrated by eople of color?? Again, the original post was about shutting down the Dan Ryan. You
5	@ChangeTheRef @delaneytarr @Alfonso_Cal @MattxRed @longlivekcx @Ryan_Deitsch @MichaelSkolnik @Alyssa_Milano @Sarahchadwickk @Emma4Change @cameron_kasky @AMarch4OurLives @JonLionFineArt
6	People with GUNS kill people these stats prove it. @NRA @MeghanMcCain @davidhogg111 @Emma4Change #NRA #GunControl #GunReform #gunviolence https://twitter.com/spectatorindex/status/1017166647478534144 …
7	@cameron_kasky @davidhogg111 @Emma4Change @SenateMajLdr @Morning_Joe @POTUS @KSibla @cowards_are_us pic.twitter.com/XbqlooDtG6
8	@Eaglewoman4 @TruthOuter Yes, at the very least, always present a balance.Though I have no faith in the generation(s) too "whatever" to show up for the first woman president, I feel a great sense of optimism for
9	.@davidhogg111 .@Emma4Change .@cameron_kasky "The Young People Will Win" https://twitter.com/selectedwisdom/status/1017139479197437952 …
10	@AMarch4OurLives @cabq @lizthomsonnm @Emma4Change @PatDavisNM I'm so sorry to have missed them! Great photos!!
11	Have y'all seen this amazing lady and her incredible talent deliver this important message?@shannonrwatts @Emma4Change @davidhogg111 @PattyArquette @RoArquette @StephenAtHome @SophiaBush @Mariska
12	@cameron_kasky @davidhogg111 @Emma4Change @SenateMajLdr @Morning_Joe @POTUS @cowards_are_us THANK YOU!!!!!!
13	Come use our facilities for free! #Roadtochange #MarchForOurLives@March4LivesSLC @davidhogg111 @Emma4Change @cameron_kasky pic.twitter.com/F4mi4ww2FN
14	Emma meets Emma! Thanks for visiting #Albuquerque @Emma4Change. @emma_hotz @katelynsarakey #MarchForOurLives #RoadToChange pic.twitter.com/wNt5Sv4l2o
15	@KevinReinholz @Emma4Change Brought tears for me too. What an incredible story.
16	@davidhogg111 @cameron_kasky @Emma4Change @AMarch4OurLives @PHXMarch4Lives @persisteresist You keep posting the same things over and over again. I've already debunked your arguments, when these so-called assault rifle bands finally make it to the Supreme Court, don't be surprised when they're struck down.
17	@Emma4Change @shannonrwatts @Stratosphantom "Hidden intention?" Her friends were killed. Put down the gun, pick up a razor.
18	@Emma4Change , @davidhogg111 , #MarchForOurLives https://twitter.com/chrismurphyct/status/1016490627742171137 …
19	@davidhogg111 @cameron_kasky @Emma4Change @AMarch4OurLives @PHXMarch4Lives @GrandvilleShow Tell Chief Justice Roberts that Justice Scalia was wrong when National Firearms Act's restrictions on machineguns were & are still constitutional. Keep proving you are too fucked up to own guns. pic.twitter.com/9Vc1kVwTMk
20	@shannonrwatts @maddsurgeon Can you not see how the Cuban flag on @Emma4Change shoulder during her speech may frighten some to some sort of hidden intention she may or may not have?

Figure 3.3: Sample tweets from dataset gathered with keywords related to parkland shooting.

Tweet No	Sample Tweets from Mid Term Elections 2018 Dataset
1	How the #WalkAway Campaign is Drowning the Blue Wave Democrats https://townhall.com/columnists/joyoverbeck/2018/07/11/how-the-walkaway-campaign-is-drowning-the-blue-
2	@TrueFactsStated What happens when your blue wave becomes a blue flush? Because that's what's coming in November. pic.twitter.com/vQz2tw0Zn2
3	We are a party of innovation. We do not reject our traditions, but we are willing to adapt to changing circumstances, when change we must. We are willing to suffer the discomfort of change in order to achieve a better future. Barbara
4	While America sleeps: Trump's treachery and the Russia scandal https://www.motherjones.com/politics/2018/07/while-america-sleeps-trumps-treachery-and-the-russia-scandal/ ...
5	@FoxNews @realDonaldTrump @real_farmacist @vandivort_david "Lulz"? Not banking on a blue wave, if it happens we are lucky! Hillary? No! In the meantime, making Dems more extreme to deal with Trump, so we can
6	@SpockResists All the news is making us apathetic. It's overwhelming. Immobilizing. We all want to protest,but even the word 'protest'has become scary and we can't seem to unite on 1 issue alone. Except #BlueWave. I propose
7	@faadiel007 feelings aren't facts and your bias is showing. Either anyone has a right to buy property anywhere, or you have group areas. Desirable suburbs attract buyers and, BoKaap will attract buyers if there are sellers.
8	@senorrinhatch Dear O,Time to take your ginkgo pills. The second civil war was won last week by the Dems. If your memory fails you every single time Trump and the GOP step on our Constitution, do not worry, come Nov 6th there
9	Good Morning @realDonaldTrump. And what is the purpose of your visit to Great Britain today; #Business, #Pleasure or Inciting #RacialHatred?#trumpUKVisit#UK#Brexit#BlueWave pic.twitter.com/3zPTzlVVKF
10	Don't complain...VOTE! Volunteer for campaigns, write postcards, make phone calls and knock on doors to get your candidates elected! If our elected officials won't change then we have to change them! #vote #bluewave
11	@MarylouRueben Marylou save yourself the hassle you will just get yourself upset arguing with them. I am on your side. We'll stick together. Blue Wave pic.twitter.com/sjNlISlX8Z
12	There's no stopping a country where truth prevails. #quote from "Veer-Zaara" film written by Aditya Chopra #indivisible #bluewave #swingleft
13	@pcdillard Yup. Appears to me a Blue Wave, if it materializes, will largely be ridden on the backs of moderates and more conventional (Kennedy/Clinton) D's. As waves usually are.
14	@1VoiceForMe @LegionDecency <Emoji: Medium star><Emoji: Flag of United States> Welcome,Brother<Emoji: Small orange diamond>WER THE RESISTANCE <Emoji: Medium star><Emoji: Beating heart>a Band of Brother's &
15	#vote #BlueWave harryshannon's photo https://instagram.com/p/BlHbB1LB1SB/
16	This is my first #FBR party! I'm encouraged by #TheResistance who want to make things better for all. #Resistance #FBRParty #BlueWave #BlueWave2018 #BlueTsunami #Resist 1. Like2. Follow3. Retweet4. Comment5. I will follow
17	#Cointelpro #Gangstalk #Gangstalking #TargetedIndividuals #Wikileak #NATO #NSA #REPUBLICANS #congress #FRABEL #DEMOCRATS #treason #TreasonousBastard #WhereAreTheChildren #BlueWave #RedWave
18	@cnnbrk Republicans petty Republicans they can sense the blueWave a coming
19	@bulldoghill Groundswell becomes Blue Wave becomes Waterworld: pic.twitter.com/KaNtSoF52K
20	@GOP what do you say to this? https://twitter.com/stonecold2050/status/1017193280339996673 ...

Figure 3.4: Sample tweets from recently dataset gathered with keywords related to the ongoing Midterm Elections 2018.

has been several protests and uproar against several groups within the US. There has been several reports that the Russian trolls are trying to spread misinformation. So we collected tweets related to the incident from dates 2018-02-14 to 2018-07-05, and with hashtags *emma4change*, *roadtochange*, *march4ourlives*, *marchforourlives*, *parkland* and *parklandshooting*. Again, these are the most trended hashtags during the time, so we have used them to collect data. We have collected in total 885,149 tweets from 311,322 users where 829,570 tweets among the total are unique. Figure 3.3 shows sample tweets that we collected related to the parkland shooting incident, using the above keywords.

For all our data collections, we scrap data from twitter using tweepy. The Tweepy, a tool that allows us to scrap the data of twitter accounts from the past as well. The advantage of using tweepy is that, unlike twitter API where we can collect data only for the past 7 days, tweepy allows us to collect data in an non-realtime fashion; data from the past can also be crawled so long as they were not removed. Since the goal of the work is to identify troll from rest of twitter accounts, the baseline must be a set of twitter accounts that have tweeted considerably about the topic of interest, the presidential elections, at the same time does not contain troll users. Other datasets as well are collected in pretty much the same way.

While experimenting, we also noticed that the Russian troll dataset, comprising of the tweets from the identified user accounts, is not gathered the same way as the others that are used in this thesis. The other three datasets —*Election 2016 dataset*, *the parkland shooting dataset* and the *midterm elections dataset* —are collected based on the then trending keywords pertaining to the respective occur-

rences, which means the tweets will be very likely relevant to the event, excluding cases where the keywords are used for some other purposes. On the other hand, the official release of the Russian troll dataset comprised of all tweets by the troll accounts over a period of time. Thus the collection tends to be more noisy than others. To counter that, we prune the dataset by removing the tweets that are irrelevant. To do that, we use the topic models learned (described in detail in Chapter 4), that discovers most relevant set of topics (each topic is comprised of a set of words) that are being discussed in the tweets. If a tweet has no overlap to any of the topics learned from all 4 datasets, then the tweet is identified to be irrelevant and removed thusly.

Also, all the tweets that used throughout the experiments, both for training and testing, are pruned to remove unnecessary words such as "http/https:", other keywords such as "rt", any stop words, any word of length lesser than or equal to 2, and punctuation and special characters.

Chapter 4

Proposed Method for Automatic Identification of Russian Troll Twitter Accounts

4.1 Proposed Approach

In this thesis, our goal is to design an automated system that learns to identify troll accounts that indulge in spreading of disinformation concerning the 2016 US presidential elections, from rest of the twitter accounts. For that purpose, we use the tweets of troll accounts identified and collected by [2], to learn a classifier that could classify troll users and tweets against tweets of a regular twitter account. In order to identify a twitter account as a *troll* account, we analyze the tweets posted and retweeted by the account over a period of time. We consider a number of

parameters including keywords that the troll account have been using and their frequency, the distribution of the type of information they have been posting over time, along with a few metadata such as the date of creation of the account, the frequency of tweets posted *etc.* to make the decision.

The objective of our research is to identify parameters that account for the behavior of the troll accounts, and use them to identify more such existing accounts in the twitter. In addition to identifying the accounts, we also attempt to analyze individual tweets and identify how similar they are to the tweets by troll accounts.

We use the collection of tweets of the troll accounts collected by [2] for our analysis. To model the baseline user behavior or the behavior of non-troll twitter accounts we use the publicly available dataset comprising the tweets about the elections from the general twitter community. Also we prune users who talk too little about politics using a set of keywords that we identify are specific to politics or about the election. Doing so allows us to identify uses who generally talk about politics, and explore questions such as how frequently a regular user talk about the election in comparison to the trolls.

We begin by introducing the methodologies used briefly followed before expanding on our proposed framework.

4.1.1 Latent Dirichlet Allocation

The Latent Dirichlet Allocation (LDA) Blei et al. [2003], is a generative probabilistic model for discovering topics from a set of documents. Given a set of textual documents (interchangeably referred to as text corpus), LDA learns to represent

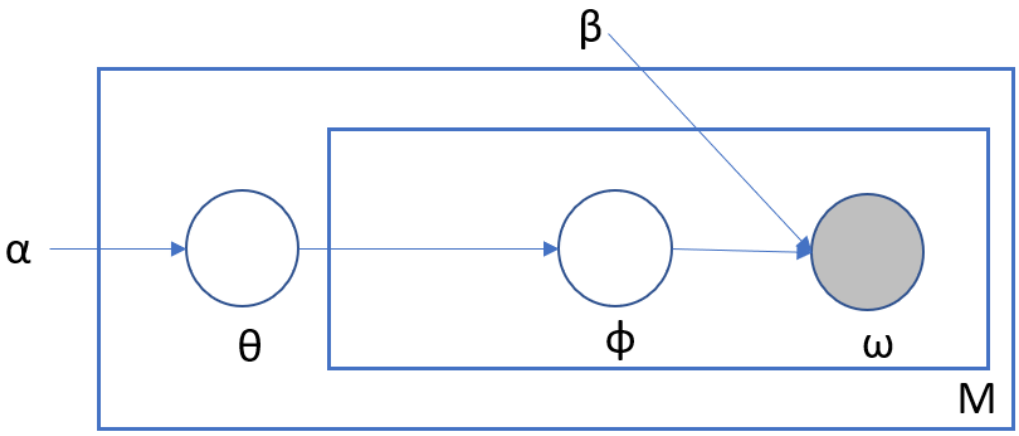

Figure 4.1: Latent Dirichlet Allocation: The outer layer represents documents and the inner layer represents the topics and words with in the topic Blei et al. [2003], where α & β are the hyperparameters.

them as a mixture of topics that outputs a sets words with the associated probabilities. The technique relies on the following assumption: when a document is created in the first place, we decide a set of topics that will represent the document, followed by writing words that explain each topic, along with a number of words for each document. The above method of writing the document can be mathematically formulated as follows: Each word in the document is (1) generated by picking the topic of interest follows a multinomial distribution (2) followed by each topic containing a set of words is in turn represented by a different multinomial distribution, where (3) the total number of words in the document is modeled by a Poisson distribution, and (4) the topic mixture is modeled by a Dirichlet distribution.

The qualitative representation of LDA is shown in figure 4.1, where, given hyperparameters α and β, LDA Blei et al. [2003] computes the joint distribution comprising of the latent variables θ, ϕ & ω where θ is the topic mixture that is modeled using a Dirichlet distribution, ϕ models the set of K topics using a multinomial distribution and ω models the set of N words using multinomial distribution. The learning is done by maximizing the likelihood of $p(\theta, \psi, \omega|\alpha, \beta)$, which algorithmically implies that given a set of documents, we learn a to identify a set of topics that the document discusses about where each topic is a set of actual words from the document that indeed makes up the topic itself, all in an unsupervised fashion.

In our framework, we use LDA to learn representative topics and keywords that makes up a normal user's twitter data over a longer time span, and compare them against a recognized troll accounts', to infer key insights into what differences they have in comparison. In other words, say an average twitter use may spend one-tenth of the tweets talking about a specific topic of interest such as the 2016 Presidential election, in contrast a troll twitter account may spend much more of their tweets in doing that. Moreover, the topics and their associated words estimated by LDA, if are significantly different between a regular and the troll users, will provide vital information about the set of topics that a troll focuses on. Further down the line, such information can help twitter to identify such troll accounts, and restrict them preemptively from spreading disinformation.

The reason behind choosing LDA as our classifier is that, it is primarily an unsupervised approach. In our dataset, despite the fact that the tweets belong to

troll accounts, not all of them are identified to be disinformation. A good proportion of the tweets do not contain information relevant to the topic of interest. Thus training a discriminative classifier such as Support Vector Machines (SVM) Cortes and Vapnik [1995] is rather tricky as the amount if noise in the data is high, discriminative classifier is not a viable option. Also SVM may not work due to another reason, twitter since not all troll accounts pertaining to the current case have been identified or removed, in there may still be more troll accounts spreading disinformation. Another advantage is that LDA is primarily a generative probabilistic model, and the size of the datasets allows for a more comprehensive modeling of topics. In addition, the representation of LDA is only partially latent unlike SVM classifiers where all representations are meant to be latent. This property comes in handy as parameters such as ω are actual words from the document and ϕ are actual topics or related set of text that talks about a specific subject. Such tangible representations of the learned model is quite advantageous for obtaining a better insight into the data as well as for the inference.

4.2 Methodology

4.2.1 Identification of Troll Users/Accounts

Given a set of tweets by a twitter account, we analyze the behavior of the user over a span of time to conclude whether the twitter account is a troll account or a regular account. In order to do that, we gather and store a collection of tweets tweeted over time by each user into a document, for both troll users and regular

users. While learning, we primarily intend to focus on the following questions

1. Is there a considerable disproportion in the keywords that these two groups use? *(Q1)*

2. How often they tweet about the topic of interest in comparison with a regular user (who also tweets about the topic)? *(Q2)*

3. Do they consistently post politically biased views? (Q3)

4. Is the metadata significantly different across the two groups, say are most accounts recently created? have their tweeting frequency increased considerably during the time of interest? *(Q4)* (for future work)

Learning

The default way to learn LDA will be to learn to unsupervisedly classify from a corpus where the LDA algorithm generates topics, in a simplistic setting, *troll* or *not a troll*. But the problem with the approach is that, such a classification is more semantic, which cannot be attributed to the LDA's topic modeling. The topic generation is modeled by learning the underlying distribution, but the distribution may not necessarily comply with the human way of interpreting the topics, in this case *troll* or *not a troll*. Due to the availability of the tweets belonging to the troll accounts, we intent to learn a independent set of topics for both troll and average twitter accounts. We compute two sets of topics, T^{tr} is a set of topics computed using the LDA from the tweets by the troll accounts. Similarly another set of topics T^{reg} are computed from the text corpus collected from the tweets

posted by the non-troll twitter accounts. Figure 4.4 shows the results of topics recovered by LDA for troll data. And and Figure 4.4 shows the topics recovered by LDA along with the set *rt*, implying the significant proportion of retweeting involved in comparison to the rest of the twitter population. The LDA models for our experiments are trained to extract 10 topics each containing 10 words, for 20 passes each, for 20 iterations, where the chunk size is set to 2000. In addition, the decay, offset and the gamma threshold are set to 0.5, 1.0 and 0.001 respectively. These parameters remain constant for all our experiments unless stated otherwise. For computing the LDA models, Genism Python library was used.

Testing

To perform testing, we use the 10% of tweets by the troll users, whose tweets were not used for training. At test time, given a set of tweets, we compute weighted scores of word overlaps along with the learned likelihoods. Topics T^{tr} is the set of topics recovered from the troll tweets via LDA, and similarly T^{reg} is recovered from the tweets of average twitter accounts, where $\{p_i^{tr}|i = 1...N\}$ & $\{p_i^{reg}|i = 1...N\}$ are the probabilities corresponding to i^{th} word of N words of each topic respectively. A tweet is comprised of a a set of words given as $y \in Y$, while a topic T is comprised of N words and there are $t \in T$ topics. Based on the words $y \in Y$ in the tweets, we computed a weighted score that represents the overlap. Each user has $|Z|$ tweets, thus we average them.

$$\mathcal{S}^{tr} = \frac{1}{Z}\sum_{z=1}^{Z}\left\{\max_{t\in T^{tr}}\left\{\sum_{i=1}^{N} p_i^{tr}(\hat{t}) * F^{T^{tr}}(t, Y)\right\}\right\} \tag{4.1}$$

where $\hat{t} = |t \cap Y|$, the Jaccard similarity is given as,

$$F(t, Y) = \frac{|t \cap Y|}{|t \cup Y|} \tag{4.2}$$

where $F_i^{T^{|*|}}$ is the Jaccard similarity (equation 4.2) the of the i^{th} word that overlaps. And similarly, the similarity between the user tweets and the LDA model learned for Election 2016 corpus is given as,

$$\mathcal{S}^{ave} = \frac{1}{Z} \sum_{z=1}^{Z} \left\{ \max_{t \in T^{tr}} \left\{ \sum_{i=1}^{N} p_i^{reg}(\hat{t}) * F^{T^{reg}}(t, Y) \right\} \right\} \tag{4.3}$$

Class C represents the class that obtained maximum likelihood among the two.

$$\mathcal{C} = max(\mathcal{S}^{tr}, \mathcal{S}^{reg}) \tag{4.4}$$

The quantitative analysis of the experiment is shown in table 4.1. From the numbers that it is evident that the Russian troll accounts are clearly differentiable from the rest as they tend to have a clear positive correlation than the other categories that discuss the same topic as the Russian trolls. We also qualitatively demonstrate the performance of the system in figures 4.9 and 4.10 with sample tweets from each user.

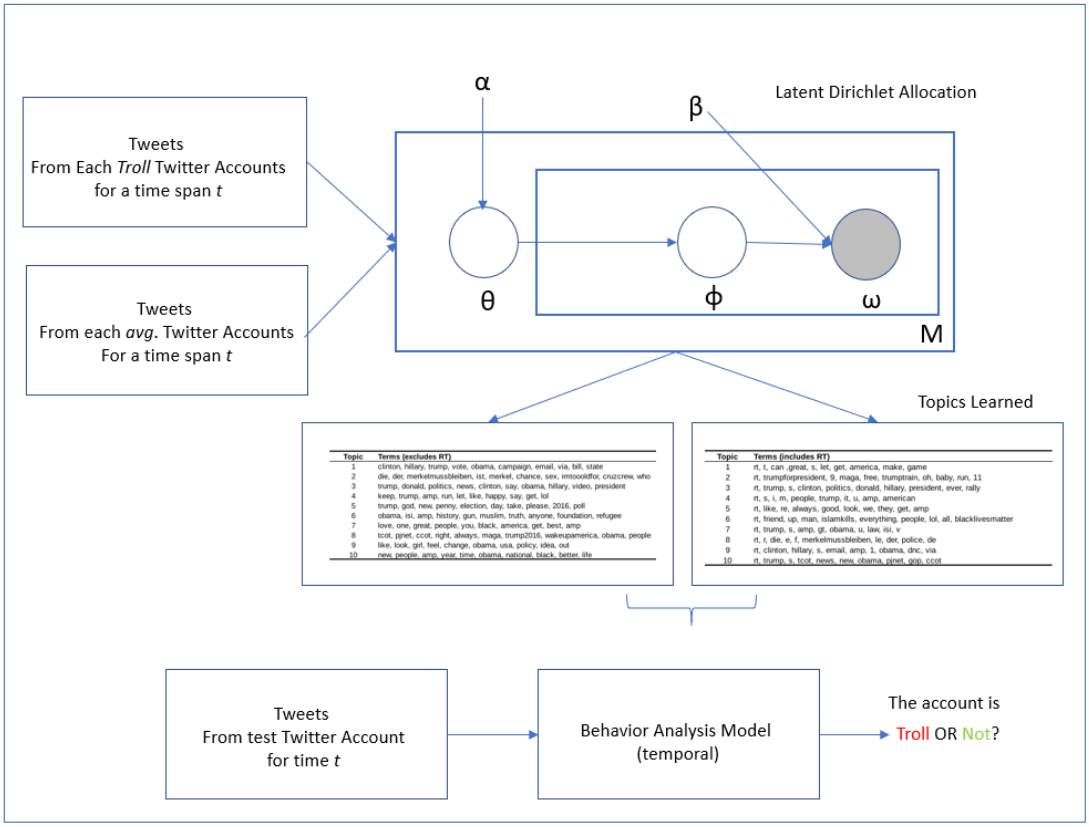

Figure 4.2: Pipeline of the proposed architecture for classification of a troll account as *troll vs. not-a-troll*.

Topic no	LDA Topics for Russian Trolls 2016 (90% of tweets)
1	trump donald news new politics post say police video election
2	trump amp want people medium hillary obama think would know
3	trumpforpresident read maga yes lol morning christmas thingsyoucantignore sta immigration
4	clinton hillary trump obama campaign bill email via state amp
5	trump obama isi trump2016 politics political john putin account used
6	black shit matter finally told get thingsmoretrustedthanhillary leave amp senate
7	tcot pjnet ccot die wakeupamerica merkelmussbleiben der merkel open star
8	trump thing great make win day right america let amp
9	white god house like people you thank got get men
10	like love one always come hate that good something people

Figure 4.3: LDA Topics computed from the training split (90% of the data) of the Russian troll dataset.

Topic	Terms (includes RT)
1	watch think right woman let muslim howtoconfuse thingsdonebymistake s auntie
2	gun rt trump 5 clinton action anti 240 28 fournette
3	rt news politics nyc co westmoreland resident in police state
4	rt trump read post new news 5 s tweet long
5	trump rt s america pum hillary causing hit strets machado
6	resolution a is what my4wordnewyearsresolution rt s text trump obama
7	rt obama s thing like trump many population closed order
8	rt hillary obama politics clinton rally people born use democrat
9	rt really s 5th school brand e378 trews russell thought
10	rt s amp trump true turn rogueone fl4trump maga happy

Figure 4.4: LDA Topics computed for Behavior analysis of troll accounts without pruning the tweets. *-*rt* to show that there is a disproportionate amount of retweets in comparison to regular twitter users.

Topic no	LDA Topics for Elections 2016 (90% of tweets)
1	nuclear state united war world power weapon tension clinton political
2	clinton trump hillary donald day ago duration email shared comment
3	trump donald campaign rally 2016 presidential republican said nominee say
4	clinton hillary state trump email secretary donald new nominee republican
5	clinton state depament foundation donor hillary people secretary press according
6	clinton hillary news trump foundation daily material mail view donald
7	location tweet always twitter learn off delete history option store
8	trump bit com hillary www right donald friend alex infowars
9	like tweet retweet aug reply liked link copy embed retweeted
10	trump clinton donald poll hillary percent show voter campaign presidential

Figure 4.5: LDA Topics computed from the training split (90% of the data) of the Election 2016 dataset.

Table 4.1: Quantitative evaluation of tweet classification: The rows represent the test set that is used for classification and the columns represents the LDA topic models learnt.

(Learned LDA Models →) (User Tweets From ↓)	Russian Trolls	Elec. 2016	Elec. 2018	Parkland Shootings
Russian Trolls	0.055235	0.035238	0.026828	0.035238

Topic	Terms (includes RT)
1	watch think right woman let muslim howtoconfuse thingsdonebymistake s auntie
2	gun rt trump 5 clinton action anti 240 28 fournette
3	rt news politics nyc co westmoreland resident in police state
4	rt trump read post new news 5 s tweet long
5	trump rt s america pum hillary causing hit strets machado
6	resolution a is what my4wordnewyearsresolution rt s text trump obama
7	rt obama s thing like trump many population closed order
8	rt hillary obama politics clinton rally people born use democrat
9	rt really s 5th school brand e378 trews russell thought
10	rt s amp trump true turn rogueone fl4trump maga happy

Figure 4.6: LDA Topics computed for Behavior analysis of the election 2016 accounts prior to pruning. *-*rt* to show that there is a disproportionate amount of retweets in comparison to regular twitter users.

Topic no	LDA Topics for Parkland shooting (90% of tweets)
1	twitter pic com roadtochange parkland host marchforourlives actor maga ted
2	florida gun violence cruz city chicago anti hot death deputy
3	parkland survivor student hogg david news nra ingraham trump massacre
4	gun control right nra law say would people conservative shot
5	medium hospital emoji sign hand skin tone light person boycott
6	emoji face state flag poker eye mark united smiling told
7	hea red heavy father emoji venezolano orange tiroteo excited del
8	parkland shooting school survivor via shooter student victim high sheriff
9	parkland roadtochange student kid people change you like vote young
10	parkland marchforourlives neveragain teen que resist allman liked thomas jamie

Figure 4.7: LDA Topics computed from the training split (90% of the data) of the gathered parkland shooting dataset.

Topic no	LDA Topics for Mid Term Elections 2018 (90% of tweets)
1	trump midtermelections bluewave vote get people like need let one
2	tear face joy life movement bailey 7780 charbonneau 0226 iphone
3	emoji water face flag state united hea sign medium skin
4	redwave maga walkaway app debatez trump2020 trump qanon free redwave2018
5	wave blue coming red november there democrat liberal gonna 2018
6	twitter pic com midtermelections california 2018 bluewave pa today closed
7	bluewave resist midtermelections trumprussia impeachtrump bluewave2018 day theresistance resistance remain
8	midtermelections democrat redwave congressional election republican 2018 dems via voter
9	midtermelections vote midterm election congress midterms2018 bluewave politics trump 2018
10	election2018 silent midtermelections via march patriot freedom redwaverising system ahead

Figure 4.8: LDA Topics computed from the training split (90% of the data) of the ongoing Election 2018 dataset.

4.2.2 Identification of Troll Tweets

In addition to classifying a twitter user into troll or not based on his activities, we dive into a slightly more challenging problem of classifying a single tweet being a troll or not. The reason why this is more challenging than identifying the troll user is that, with the latter we have a rich temporal information that allows us to model a slightly more complex behavioral model, that can keep track of the user's activities over time, have access to the time-progressed data associated with the user account. But in this problem, we use only a single tweet of any random user to score how likely it is to be a troll account.

While the previous model has a sophisticated user behavior-based reasoning, it becomes a real issue when the troll users, to avoid from being spotted, try intentionally to behave more like regular users; post tweets about various general topics, tweets that does not represent biased views of theirs, in order to confuse the detection systems. In such cases, it becomes a real challenge, as they can easily

Tweets from user 'stlouisonline'			
Russian Troll	**Elections 2016**	**Midterm elections**	**Parkland shooting**
0.04874996	0.02815131	0.024245268	0.02815131
Sample Tweets			
Convention update: Al Franken turns to comedy to skewer Trump #politics			
Analysis: Trump won't change; he can't let go of a grudge https://t.co/57MPriURba			
Trump could have avoided paying federal taxes for 18 years, report says https://t.co/uKxFoUYhV2			
Franks wins big in re-do election for 78th District state representative seat #StLouis			
Judge orders new election for Hubbard/Franks state representative race in St. Louis #politics			

Tweets from user 'robertebonyking'			
Russian Troll	**Elections 2016**	**Midterm elections**	**Parkland shooting**
0.05051941	0.028476125	0.02593657	0.028476125
Sample Tweets			
RT @TheDailyEdge: Not sufficiently impressed w/the dignity & poise Obama has displayed consistently for			
RT @AnandWrites: Why is this overt racism not enough for more people to be braver? https://t.co/2yc5Y2THb(			
RT @LorenzaOrrin: Def ready to check it out https://t.co/JLEEc9WOkY			
RT @thehill: Humane Society launches ad: Trump presidency a 'threat to animals everywhere' | WATCH: https:/			
RT @Yamiche: This Data Download on @meetthepress segment is key. Voters are basically as concerned about			

Tweets from user 'willisbonnerr'			
Russian Troll	**Elections 2016**	**Midterm elections**	**Parkland shooting**
0.04930852	0.027206434	0.025694303	0.027206434
Sample Tweets			
RT @oneunited: Challenge yourself this #NewYear to move your finances to a Black owned bank. Take the #Bar			
RT @duttypaul: RRR!!! DO U REMEMBER DIS 1?!?!?!?! #DUTTYSTEPPINZ RRR!!! SWIPE IT!!! https://t.co/U5vMT			
RT @Welcometoharlem: #Quote - Start where you are. Use what you have. Do what you can.....Arthur Ashe htt			
RT @rapstationradio: #NowPlaying: "Customer Service" Jurassic 5 - https://t.co/LBtwx7rpdC https://t.co/pMah			
RT @kwamerose: "Can Stein Capitalize on Alienated Sanders Delegates" My Exclusive 1on1 w/ Green Party Can			

Tweets from user 'lazykstafford'			
Russian Troll	**Elections 2016**	**Midterm elections**	**Parkland shooting**
0.04814282	0.0274817	0.02378609	0.0274817
Sample Tweets			
Hillary isnâ€™t a good senator so she canâ€™t be a good president #HillaryNoThnx			
RT @summersstine: Hosted by POS #LindaSarsour herself https://t.co/d2xGBk8Bqe			
RT @adriennefunny: @tonyposnanski Ha ha ha.... .Hello Tony!!! Happy Holidays!			
RT @prd_2b_USA: I wish there was just as much fanfare when members of #Congress replace many #dumbass			
Don't go into business to get rich. Do it to enrich people. It will come back to you.			

Figure 4.9: Sample users and their tweets from the Russian Troll dataset with their corresponding similarity scores computed for each of the learned model.

Tweets from user 'youjustctrlc'			
Russian Troll	**Elections 2016**	**Midterm elections**	**Parkland shooting**
0.050448284	0.028367074	0.025878072	0.028367074
Sample Tweets			
RT @DebFischerNE: It would be wise for him to step aside and allow Mike Pence to serve as our party's nomine			
RT @daniecal: I really hate how this woman is being used https://t.co/oPTpS7FWKo			
RT @KeshRue: We have had enough https://t.co/7UszkAUWdH			
RT @AriMelber: Wow former Trump Hispanic adviser Jacob Monty, who resigned, says it's now clear Trump "dc			
RT @ABC7: #LIVE Trump protesters estimated to be 10,000 says LAPD https://t.co/u5lTveArpy https://t.co/7n9			

Tweets from user 'blmsoldier'			
Russian Troll	**Elections 2016**	**Midterm elections**	**Parkland shooting**
0.046096936	0.026566919	0.022902206	0.026566919
Sample Tweets			
RT @eclecticbrotha: Tune in tomorrow when we try to humanize a racist sex abusing sociopath https://t.co/YCf			
RT @Yamiche: This Data Download on @meetthepress segment is key. Voters are basically as concerned about			
RT @JordanChariton: .@HillaryClinton camp wrote gun hit piece on @BernieSanders; then put gun violence vict			
RT @mrtstur: Busted... @BlogLiberally https://t.co/p3hFcc7IF9			
RT @Omojuwa: US markets not fully buying a ClintonÂ win. https://t.co/OIeqSDabGk https://t.co/PEKzdn2FRm			

Tweets from user 'glennharper_'			
Russian Troll	**Elections 2016**	**Midterm elections**	**Parkland shooting**
0.07164621	0.04953695	0.040447753	0.04953695
Sample Tweets			
#Guns4NY When seconds count, the police arrive in minutes. Or hours			
#demndebate Probably, weâ€™ll get out of national debt if we stop wasting money? Donâ€™t you think so? #D			
Only one thought #Prayers4California https://t.co/aGoMJeGFdv			
It would not be so much noise on the matter if it was the first time #Prayers4California			
I don't carry a gun because I'm paranoid			

Tweets from user 'dannythehappies'			
Russian Troll	**Elections 2016**	**Midterm elections**	**Parkland shooting**
0.05354066	0.035556797	0.027647214	0.035556797
Sample Tweets			
RT @Sttbs73: #RejectedDebateTopics Which is your favorite groping hand Mr Trump?			
#IHatePokemonGoBecause they keep patching my hax.			
RT @ATHEIST_Blessed: Everything Everyone Says Proven Wrong; Only Trump Correct. #TrumpsFavoriteHeadlin			
My #ChildrenThinkThat Obama is scary			
RT @LoLo_OOC: #TrumpsFavoriteHeadline Donald Trump is now considered a gender			

Figure 4.10: Sample users and their tweets from the Russian Troll dataset with their corresponding similarity scores computed for each of the learned model.

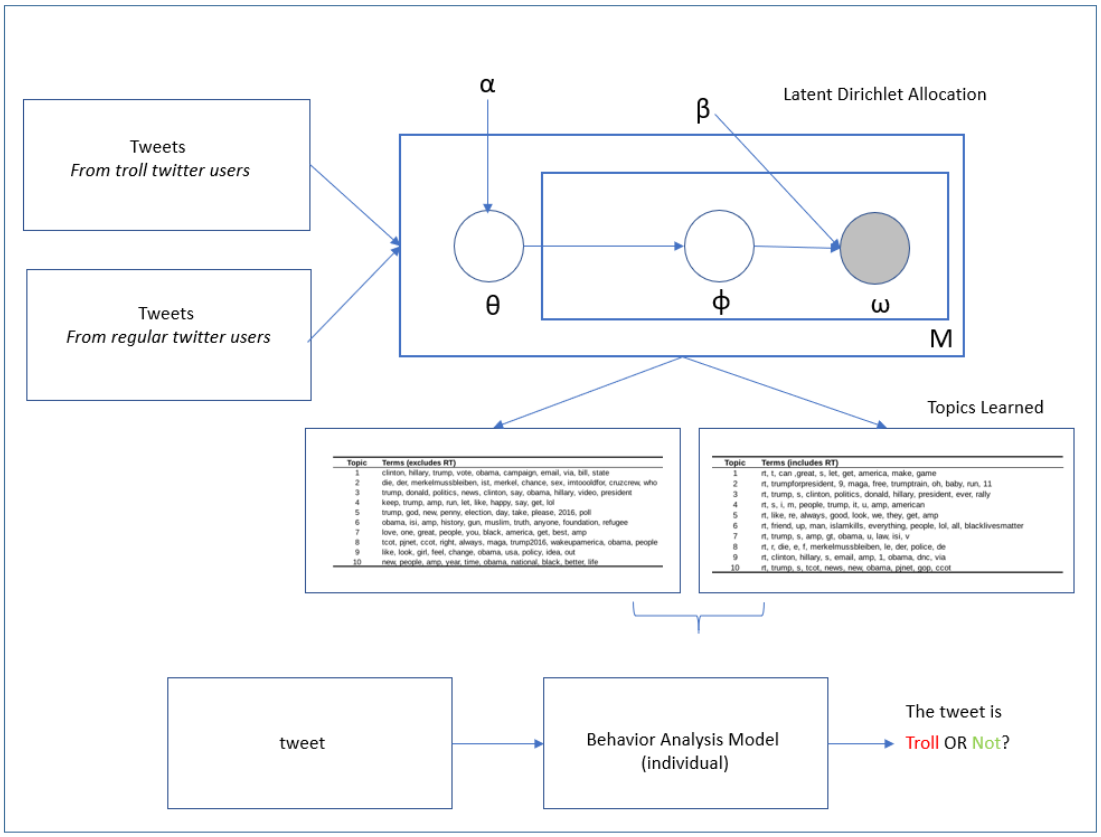

Figure 4.11: Pipeline of the proposed architecture for classification of a single tweet as *troll vs. not-a-troll*.

go undetected. It is very likely that such troll accounts pertaining to the case of interest in this paper, are certainly out there, continuing to spread misinformation. In order to tackle this we extend our classification method to classify single tweet. Given a tweet, our method computes a score for how likely it is being tweeted by a troll account, based on the similarity in the textual space, between the tweet and learned topics, considering word overlap as well as the likelihood of the overlapping word with respect to the topic.

Learning

Similar to 4.2.1, we primarily learn two different topic models, T^{tr} is a set of topics computed using the LDA from the tweets by the troll accounts where another set of topics T^{reg} are computed from the text corpus comprising tweets gathered from regular twitter accounts who discussed the election 2016. Unlike the previous section, where we split the train and test sets using the user IDs, in this section, as the goal is to classify individual tweets, we split the tweets randomly without considering the user IDs. Figure 4.4 shows the learned topic model using LDA for the troll data, and also Figure 4.4 shows the topics recovered along with the set *rt*, implying the significant proportion of retweeting involved in comparison to the rest of the twitter population that talked about election as shown in Figures 4.4 and 4.4. In addition to differences in the topics that the two groups discuss, it is also evident that the number of retweets among the two populations are disproportionate; the Russian troll accounts retweet much more than the regular twitter users that tweeted regarding the elections.

4.3 Results and Evaluation

Unlike 4.2.1, at test time, in this section we deal with a single tweet than a set of tweets of an user collected over time. Given a tweet, we perform filtering as discussed in 3, then compute a Jaccard similarity measure that embeds word matching scores and the likelihoods learned via LDA. Topics T^{tr} is the set of topics recovered from the troll tweets, and T^{reg} is recovered from the tweets of regular twitter accounts, where $\{p_i^{tr}|i = 1...N\}$ & $\{p_i^{reg}|i = 1...N\}$ is the probability corresponding to i^{th} word of N words of each topic respectively. A tweet is comprised of a a set of words given as $y \in Y$, while a topic T is comprised of N words and there are $t \in T$ topics. Based on the words in the tweets, we computed a weighted score that represents the overlap

$$\mathcal{S}^{tr} = \max_{t \in T^{tr}} \left\{ \sum_{i=1}^{N} p_i^{tr}(\hat{t}) * F^{T^{tr}}(t, Y) \right\} \tag{4.5}$$

where $\hat{t} = |t \cap Y|$ and $F_i^{T^*}$ is the Jaccard similarity the of the i^{th} word that overlaps. And similarly, the similarity between the tweet and the LDA model learned for Election 2016 corpus is given as,

$$\mathcal{S}^{reg} = \max_{t \in T^{tr}} \left\{ \sum_{i=1}^{N} p_i^{reg}(\hat{t}) * F^{T^{reg}}(t, Y) \right\} \tag{4.6}$$

Class C represents the class that obtained maximum score among the two.

$$\mathcal{C} = max(\mathcal{S}^{tr}, \mathcal{S}^{reg}) \tag{4.7}$$

The quantitative analysis of the experiment is shown in tables 4.2 & 4.3. The aim of the analysis is to show that the Russian trolls have a specific behavior that can be contrasted from the rest of the twitter users, given they discuss the same topic, the US presidential elections 2016. The table 4.2 shows that the similarity between the test tweets that belong to Russian Trolls (ground truth) and the Model learned from Russian Trolls is significantly higher compared to the general twitter community that discussed the same topic ($0.0469 > 0.0286$). Likewise, the tweets of the general twitter community and the Russian trolls is significantly dissimilar ($0.0629 < 0.1171$). Given a general twitter community's discussions and a Russian troll's tweets (identified by authentic sources), we are able to isolate them and show that the dissimilarity between them is not marginal, but rather considerable. In other words, the magnitude of differences demonstrate that the dissimilarity arises not simply from the difference in the data, but from the nature of the users themselves. This is further strengthened by extending the analysis to compare with the tweets pertaining to the recent occurrence of the Parkland Shootings and the ongoing Midterm Elections as well. It is evident that the degree of dissimilarity is considerably high. Also we demonstrate our results on the pruned Russian Troll dataset; the Russian troll dataset comprises all tweets which may not be relevant to the 2016 election whereas the other datasets is made of tweets that belong to the respective events. Thus we pruned the dataset as discussed in 3 to remove irrelevant tweets. The results are even better as the extent of dissimilarity is further increased 4.3. Also in addition to the quantitative evaluation, qualitative evaluations are included where we show the tweets and their corresponding scores

Table 4.2: Quantitative evaluation of tweet classification: The rows represent the test set that is used for classification and the columns represents the LDA topic models learnt.

(LDA Models →) (Tweets From ↓)	Russian Trolls	Elec. 2016	Elec. 2018	Parkland Shootings
Russian Trolls	0.0469	0.0286	0.0232	0.0201
Elections 2016	0.0629	0.1171	0.0337	0.0354
Midterm Elections	0.0330	0.0258	0.1044	0.0539
Parkland Shootings	0.0340	0.0214	0.0437	0.0848

w.r.t. each learned LDA model in figures 4.12, 4.13, 4.14 and 4.15.

In the qualitative results, we can observe that the values at times are close especially for example in Figure 4.12. The reason for such closer values is that, the similarity measure is based on a few factors, the most important one is the length of the tweet. The varying lengths of the tweets constitutes to the inconsistency in the scale of the similarity, thus the variance in the similarities. In the qualitative results, it has to be noted that with tweets that contain words such as "fake" and other biased statements are evidently categorized as trolls. Also to populate the tables in the qualitative results, we randomly chose tweets from each dataset than cherry pick them, for fair comparison, thus the results are not all in favor. But the quantitative results suggests that the overall classification is in agreement with our hypothesis.

Tweets from Russian Troll Dataset	Russian Troll	Elec 2016	Elec 2018	Parkland Shooting
RT @Conservatexian: News post: "TWITTER Buries 32 of Donald TrumpÂ's GOTO Battleground Tweets! Â…Deletes Another!" https://t.co/XZLachsMsL	0.23529412	0.1666667	0.05	0.10526316
Who could be on the stump for Hillary Clinton and Donald Trump in Ohio: Ohio Politics Roundup: Who's stumping fo... #Cleveland #politics	0.14285715	0.2	0.09090909	0.04347826
RT @redsforblacks: "I plan to vote for Donald Trump OR Hillary Clinton." #MakeMeHateYouInOnePhrase	0.21428572	0.3076923	0.13333334	0.0625
RT @TheFinalCall: Hillary Clinton and Donald Trump: Which one is worse: Lucifer, Satan, or The Devil? [WATCH] https://t.co/iCnBGznsBvâ€¦	0.1764706	0.25	0.11111111	0.05263158
RT @zeynep: Pre-election from NYT, WaPo & Politico: Clinton email server (first) outnumbered Trump conflict-of-interest stories five to oneâ€¦	0.17391305	0.125	0.08	0.03846154
RT @vannsmole: Donald Trump Sends Hillary Clinton a â€˜Get Wellâ€™ Message for Her Pneumonia https://t.co/9oTRbMeqsO	0.1875	0.2666667	0.11764706	0.055555556
RT @BooyahBoyz: Donald Trumpâ€™s OUTRAGEOUS? Some say He's a CLOWN BUT He says the things That NEED 2 be said As OUR COUNTRYâ€™s GOIN DOWN #Truâ€¦	0.15789473	0.2222222	0.1	0.04761905
A Trump election could harm L.A.'s Olympics bid, Mayor Garcetti says https://t.co/zlqydkkvrO #politics	0.25	0.1111111	0.1764706	0.05263158
Why Donald Trump might not debate Hillary Clinton #politics	0.21428572	0.3076923	0.13333334	0.0625
Two new polls suggest tight race in Ohio between Donald Trump and Hillary Clinton https://t.co/UXHOatyLFV #politics	0.22222222	0.2941177	0.1	0.04761905
The Media wants Trump supporters 2 despair & not bother 2 vote. Do NOT believe the FAKE exit polls ðŸ‘‰ get out and VOâ€¦ https://t.co/Xdp8l0OR0O	0.21052632	0.0952381	0.15	0.045454547
RT @Conservatexian: News post: "Kellyanne Conway: I could work outside administration to 'haunt' Donald Trump opponents" https://t.co/AY5NEâ€¦	0.22222222	0.1578947	0.04761905	0.1
Hillary Clinton, Donald Trump discuss plans to combat terrorism after bombings: Politics Extra https://t.co/7wnQR2H6zT #politics	0.15789473	0.2222222	0.1	0.04761905

Figure 4.12: Sample tweets from the Russian troll dataset with their corresponding similarity scores computed for each of the learned model (w.r.t. each of the four category/datasets on which models are learned on)

Tweets from Elections 2016 Dataset	Russian Troll	lections 201	dterm Electic	Parkland Shooting
CNN has decided that the letter from the Democrat nomineeâ€™s doctor â€œdebunksâ€ worries over her health but at the same time the cable network insisted that a similar letter...	0.05970149	0.0923077	0.02898551	0.014285714
Hillarys fundraiser host reported to compare Trump to Hitler and other dictators her whole speech...	0.02197802	0.1204819	0.03333334	0.010869565
I just didnâ€™t know when but it seems the Democratic Partyâ€™s white leadership has given the Congressional Black Caucus their talking points as they are out in force trying to slap down Donald Trump as a racist who doesnâ€™t care about black people...	0.05714286	0.0277778	0.04225352	0.01369863
Now Tied 0 Comments Daily Beast: â€˜No One Was Shockedâ€™ That CNN Cancelled Dr. Drew Pinsky 0 Comments Paul Nehlen: Hillary Clinton Can Prove Sheâ€™s Opposed to TPP by Urging Dems to Block Lame-Duck Passage...	0.06060606	0.0769231	0.02941177	0.014492754
Pittsburgh Tribune-Review editorial page columnist Salena Zito called in to discuss her latest piece in the New York Post. Donald Trumpâ€™s support is coming from middle-class people who worry their country is going down the tubes â€“ and is taking their children	0.11764706	0.0857143	0.05555556	0.027027028
Right? Why Are You Even Here?â€™ - POLITICO Magazine Shared 6 days ago (AUDIO) HILLARY CLINTON FULL INTERVIEW WITH ANDERSON COOPER...	0.05172414	0.1296296	0.01666667	0.033898305
according to a new NBC News/SurveyMonkey Weekly Election Online Tracking Poll. Clinton is favored by 50 percent of registered voters	0.13636364	0.1904762	0.08695652	0.041666668
by trying to tie the Ku Klux Klan and other hate groups to Donald Trump and his historic run for President. As an African-American and an enthusiastic supporter of Trump I am deeply...	0.07272727	0.0925926	0.03508772	0.01724138
including an Arabic religious TV network with a direct tie to Hillary Clintonâ€™s top aide Huma Abedin. Both Prince Alwaweed Bin Talal and Murdochâ€™s Fox News network...	0.08333334	0.1470588	0.02631579	0.054054055
...call for a special prosecutor to investigate her. (AP Photo/Gerald Herbert) Trump manager: FBI director may 'change his mind' on Clinton By (@anna_giaritelli) ...	0.07843138	0.1	0.01851852	0.03773585
the New York Times reported that Clinton told FBI officials former Secretary of State Colin Powell had advised her to use a personal email account while she held the Secretary of State office herself. â€œHer people have been trying to pin it on me	0.0882353	0.15625	0.02777778	0.057142857

Figure 4.13: Sample tweets from the Election 2016 dataset with their corresponding similarity scores computed for each of the learned model.

Tweets from Parkland Shooting Dataset	Russian Troll	lections 201	dterm Electic	Parkland Shooting
@ShoutAloudNow OH LOOK! Even more GOP raping, pillagingWe need these privatization GOP's Scams Right?#TYTLIVE @cenkuygur @RealTimBlack @LeeCamp @dpakman @Amy_Siskind @DanRather @krassenstein #WednesdayMotivation #Resistors...	0.02439024	0.05	0.13513513	0.07692308
Yet @SenateMajLd & @SpeakerRyan are<Emoji: Speak-no-evil monkey><Emoji: Hear-no-evil monkey><Emoji: See-no-evil monkey>to this serious threat to<Emoji: Flag of United States><Emoji: Exclamation question mark>...	0.03333334	0.0508475	0.06896552	0.0877193
@SenMcCain @JeffFlake<Emoji: Eyes><Emoji: Right pointing backhand index>"evil"@realDonaldTrump represents to<Emoji: Flag of United States>Democracy,& @SenateMajLdr @SpeakerRyan also KNOW the danger of this @POTUS...	0.02083333	0.0652174	0.08888889	0.13953489
Excellent Strategy from @MalcolmNance <Emoji: White heavy check mark> Can we mobilize 20 mil or more<Emoji: Exclamation question mark> I say YASSS<Emoji: Hugging face> We need a few org to get to work on this...	0.02298851	0.0113636	0.03488372	0.047058824
@Emma4Change If people don't think young lives are enough to prompt the change you want to see, let them know about the NRA-Russia connection in the Meuller investigation? ..	0.14814815	0.0333333	0.06896552	0.10714286
@Emma4Change - The country NEEDS your voice in order to motivate young people to turn out and VOTE in November. ..	0.06896552	0.0333333	0.10714286	0.14814815
The world discourages me, but people like @MomsDemand @rgay and others fighting for their country inspire. ...	0.06896552	0.0333333	0.14814815	0.10714286
@KyleKashuv @fred_guttenberg Is like talking to a wall, donâ€™t let those bullies get to you & your family ...	0.05555556	0.027027	0.08571429	0.22580644
@Emma4Change @PatsGirlUSA BADASS LEADER.....WTGG!!!!<Emoji: Person raising both hands in celebration><Emoji: Person raising both hands in celebration><Emoji: Person raising both hands in celebration>...	0.00363636	0.0072993	0.02222222	0.026022306
I wanna personally give a huge thanks to @AMarch4OurLives for going on this #RoadToChange . I've never felt so inspired to know that people like @Emma4Change...	0.0625	0.030303	0.09677419	0.13333334
@davidhogg111 @Emma4Change @elee1025 @AnnMarieMooney2 @amandablount2 @stacyanngoodman @NYCanegirl @JonC54 Thatâ€™s right....	0.11111111	0.0344828	0.2	0.25
#DonTheCon #Isupportmueller Americanâ€™s 4 Change <Emoji: Flag of United States>REGISTER TO VOTE...	0.02272727	0.0465116	0.09756097	0.125
We met and hugged @Emma4Change, hugged @cameron_kasky, wanted to hug @davidhogg111 but that young man was ALWAYS engaged in conversation. You all are truly inspirational. #RoadToChange pic.twitter.com/AQS8KcnPWB	0.03703704	0.0769231	0.12	0.16666667

Figure 4.14: Sample tweets from the parkland shooting dataset with their corresponding similarity scores computed for each of the learned model.

Tweets from Elections 2018 Dataset	Russian Troll	lections 201	dterm Electic	Parkland Shooting
@1VoiceForMe @LegionDecency <Emoji: Medium star><Emoji: Flag of United States> Welcome,Brother<Emoji: Small orange diamond>WER THE RESISTANCE…	0.00787402	0.015873	0.05785124	0.032258064
@StormyDaniels @JimMallett4 @MichaelAvenatti BLUEWAVE RESISTANCE United WE Stand. Theyâ€™ll keep coming at us, and weâ€™ll fight. Weâ€™re at war…	0.02380952	0.075	0.13157895	0.102564104
@TomSteyer @abbytencat <Emoji: Water wave><Emoji: Water wave><Emoji: Water wave>1/2We MUST UNITE as "People Against Trump" (ANY party) But have 2 #VoteBlue…	0.07317073	0.0232558	0.12820514	0.04761905
@StephenKing @tedcruz @MikeDiaz285 @CARLENEGE123 @NiceHarley Apparently your alt fake news has not been keeping you informed…	0.06060606	0.0294118	0.12903225	0.09375
@seanhannity Bannon is wrong (I <Emoji: Heavy red heart> saying that) impeachment isnâ€™t going to end the #BlueWave…	0.05555556	0.027027	0.08571429	0.11764706
@TheRickyDavila @mandyevans <Emoji: Water wave><Emoji: Water wave><Emoji: Water wave><Emoji: Water wave><Emoji: Water wave>…	0.00578035	0.0116279	0.04819277	0.0295858
Coming Summer 2018! "In The Know With Wala Blegay", delves into the national and local political issues that are affecting our communities…	0.05882353	0.0285714	0.16129032	0.09090909
<Emoji: Medium star><Emoji: Flag of United States>Go! BETO!<Emoji: Beating heart>KICK ASS!<Emoji: Medium star><Emoji: Raised fist (medium dark skin tone)>!<Emoji: Raised fist (light skin tone)>!…	0.00699301	0.0140845	0.05882353	0.035971224
VOTE VOTE VOTE VOTE VOTE BLUE WAVE<Emoji: Water wave><Emoji: Blue heart><Emoji: Flag of United States> https://twitter.com/costareports/status/101713910051057254 6Â â€¦	0.03846154	0.08	0.2857143	0.17391305
#NATOSummit2018 Trump is destroying our country. We miss President Obama. Blue Wave is coming <Emoji: Water wave>….	0.07142858	0.0465116	0.15384616	0.09756097
Challenge Accepted. #gameON #FixIt #BlueWave <Emoji: Flag of United States><Emoji: Victory hand (medium light skin tone)><Emoji: Heavy red heart> https://twitter.com/realdonaldtrump/status/101715381362562662 4Â â€¦	0.03448276	0.0714286	0.3043478	0.25
â¦@GOPâ© Do not underestimate us. #MeToo, #TimesUp, and millions of women who have no intention of losing their rights, will be voting in November. #VoteThemAllOut …	0.03333334	0.0689655	0.14814815	0.10714286
#FBR partyLIKE<Emoji: Heavy red heart>RT<Emoji: Universal recycling symbol>FOLLOW ALL WHO RT<Emoji: Water wave>#TheResistance #Resistance #ImpeachTrump …	0.02325581	0.0476191	0.12820514	0.1

Figure 4.15: Sample tweets from the Election 2018 dataset with their corresponding similarity scores computed for each of the learned model.

Table 4.3: Quantitative evaluation of tweet classification on pruned Russian troll dataset: The rows represent the test set that is used for classification and the columns represents the LDA topic models learnt.

(LDA Models →) (Tweets From ↓)	Russian Trolls	Elec. 2016	Elec. 2018	Parkland Shootings
Russian Trolls	0.0884	0.0565	0.0440	0.0377
Elections 2016	0.0677	0.1260	0.0366	0.0381
Midterm Elections	0.0556	0.0348	0.1065	0.0553
Parkland Shootings	0.0567	0.0287	0.0480	0.0917

Chapter 5

Further Extension to Encode Semantic similarity

While the Latent Dirichlet Annotation (LDA) Blei et al. [2003] in chapter 4 provides us a consensus of words as topics, which enables us to match words and compute a similarity score for a tweet or even a user behavior for over a period of time. At the same time it has to be noted that, they only allow us for direct matching of words and ignores the underlying semantic similarity between the words or sets of words. Capturing semantic similarity would imply that the topics and tweets match semantically than just the word similarity. But what are the advantage of semantic similarity? Given a pair of set of words, Humans infer higher level contextual similarity and be able to derive a more meaningful information, and not merely a measure of overlap. Lets take a look at a famous example used in the Natural Language Processing (NLP) community. A king to

a man is what a woman to a queen. So, assuming we can represent each word in a hypothetical space, where each word will be denoted by a vector, the distance between the word *King* and the word *Queen* and the distance between the words *man* and *woman* should be approximately the same. And King - Man + Woman = Queen. Such semantic similarity measure has been observed with the recent word2vec Mikolov et al. [2013] models. Word2vec is a tool that implements the continuous bag-of-words and skip-gram architectures that allows the words be represented as vectors. The word2vec is learned to map each word to a hypothetical space as embeddings in a linguistic-context preserving manner, thus at test time, given a word they generate a vector representation. More details about the semantic properties, with examples, can be found at Mikolov et al. [2013] and [3].

For our experiment, we intend to leverage the semantic properties via word embeddings and exploit the similarities/dissimilarities between them to reason the proximity of a tweet or a series of user tweets to belong to Russian troll or not. The word2vec is a shallow 2-layer neural networks that groups semantically similar words into proximal regions in the word embedding space. The concept of synonymy is represented as a distance measure in the continuous space. For a given word in a tweet, we compute the distance between the word and the each word of each topic, and an average of distances is then used to choose the topic to which the tweet belongs to. We use cosine distances as they capture the vector distance thus a better measure of dissimilarity especially in the vector space. In addition they are normalized by nature and does not require additional normalization, thus is easily comparable to distances between any other pair of word vectors.

5.1 Methodology

In order to compute the similarity between tweets and learned topics, through the word2vec space, we take a slightly altered approach. Given a set of topics T learned via LDA, and the tweet Y containing y words, the similarity is given as,

$$\mathcal{S}^{tr} = \frac{1}{Z} \sum_{z=1}^{Z} \left\{ \max_{t \in T^{tr}} \left\{ D^{T^{tr}}(t, Y) \right\} \right\} \tag{5.1}$$

where $|Z|$ is the number of tweets. Also in the same way S^{reg} is calculated as described in sections above. The formulation in equation 5.1, remains the same for most part except here the distance is not similarity or overlap of the words, rather a more sophisticated distance in a vector space learned via word2vec. To compute the vector similarity we use cosine distance metric 5.2 which can be given as,

$$D^{T^{tr}} = \frac{t \cdot Y}{||t||\ ||Y||} \tag{5.2}$$

The primary reason behind the use of cosine similarity is that, other distances such as euclidean distances captures the distances between magnitudes whereas the cosine distances calculates the cosine angle between the vectors. Thus the metric quantifies orientation of vectors in the vector space rather than differences in magnitudes.

Table 5.1: Quantitative evaluation of tweet classification via semantic similarity (using Word2vec) on the Russian troll dataset: The rows represent the test set that is used for classification and the columns represents the LDA topic models learnt.

(LDA Models →) (Tweets From ↓)	Russian Trolls	Elec. 2016	Elec. 2018	Parkland Shootings
Russian Trolls	0.1678907	0.132171	0.152742	0.129949

5.2 Results and Evaluation

Table 5.1 shows quantitative evaluation of the word2vec based similarity estimation for classifying tweet. The results are quite in correlation with the ones we had earlier in chapter 4. As mentioned in the above sections, we hypothesize that the behavior of the troll users are differentiable from the rest of the twitter community given they tweet about the same events over time, thus allows us to categorize a new user or even a tweet to belong to one of the categories. The numbers in table 5.1 strengthens our argument, as it is evident that the troll accounts are more easily identifiable and exhibit a strong positive correlation toward the topic model learned form the troll data ($0.1678 > 0.1321$), than other topic models which are of relevance; especially the Election 2016 where the topic of interest (of both parties, troll and regular tweeps) is quite the same.

Chapter 6

Chinese 50-cent army

Another popular incident of this sort is the *Chinese 50-cent army*, a colloquial term for such players that were hired by the Chinese Authorities in an attempt to manipulate the public opinion to be in favor of the Communist party of China [8]. Unlike the trolls that were used in the 2016 US Elections, these commentators write contents that are in a way propaganda of the Communist party of China. While that appears normal, the issue is that the players have been indulging in a massive spreading of misinformation. Also according to [7], the propaganda is state funded and these commentators are hired and paid for by Chinese authorities. It has also been believed that the players had been paid 50-cent each for doing so, thus the name 50-cent army[10]; a claim which is also argued to be incorrect by King et al. [2017]. Also some party members have been used to spread the propaganda, as a part of their party duties. The main goal of these 50-cent army "officers" is to write comments favorable and also derail discussions that are unfavorable to the

Communist party [9]. They also have spread narratives in favor of the ruling governments and write derogatory comments against their political opponents to undermine the opponents and critics [9].

King et al. [2017] claims that, on contrary for the 50-cent army to be a state sponsored propaganda, most commentators were paid for by the bureaucrats who were merely responding to the government by spreading pro-government talks by doing so, during the times of crisis. King et al. [2017] also found that most of the commentators (80% of them) were merely cheerleading for the government with inspirational slogans and around 13% indulging in praising the government and its policies, and have rarely engaged in direct arguments. Later on in December 2014, an archive of the emails to and from the account that is responsible for the internet propaganda, is released by a anonymous blogger (Xiaolan xiaolan.me/). The administrative account from which the transactions have been released, belong to the propaganda department of Zhanggong district of Jiangxi province, a country-level administrative unit. The archive from the email had contents ranging from reporting activities of the 50-cent army to claiming credit for completing the commentators' assignments among various other communications. While the reports and journals have managed to peek into the data, by reporting a few important documents, the archive by itself, is quite large and contains data of diverse types including raw text, text from other emails, screenshots, along with word and excel formatted attachments. Such diverse formats of data along with several other links to external information has deemed impossible to consume the data even by some of the most advanced available text analysis techniques. Thus no systematic and

complete interpretation has been done to the best of our knowledge on the data when our team began the work.

For our analysis of the 50-cent army, we identified 2,341 emails sent between 2/11/2013 and 11/28/2014. Of the identified emails, 1,208 contained usually multiple 50-cent posts; easily processable by text processing methods. Overall, we extracted 43,757 known 50-cent posts from these emails and their respective attachments, that form a basis for our analyses where we use them as our training set, and learn models that help identify other known 50c posts, and potentially be able to identify more that are to be posted - a generalized model that can identify such propaganda posts in the future.

The goal of the project was to learn if the 50-cent army had known any comments and use them repeatedly or across users, as such similarity would strengthen the argument that they were a group of organized servants that indulged in propaganda of misinformation. The scope of my work in the project, for most part, is limited to gathering data and doing primitive analysis, where I collected data from three government sites [11] for over a period of time. The problem is currently being pursued by another member of our team. The data is collected in a similar way as detailed in Chapter 3, where we identified keywords that the users are known to use via the released email dataset. One of the major issue that we faced was that, several comments were getting censored within a few hours of publication, thus the data collection has to be continuous and seamless in order to gather data without much loss.

While my primary contribution to the project is with data with minimal anal-

ysis, the method devised and demonstrated above is almost directly applicable to the dataset and the problem as well; to analyze user behavior, analyze comments and categorize them.

Chapter 7

Conclusion

We have proposed a framework that successfully demonstrates its ability to identify troll accounts from a regular twitter account. The thesis addressed one of the most crucial problems affecting the information space by propagating disinformation. As promised the proposed approach performs considerably well, in identifying the troll accounts. We also proposed a troll tweet identification, that addresses the same problem but instead of learning sophisticated models that reasons the user behavior over time, the approach focuses only on the tweet to identify troll user. The goal is not to identify but rather provide a predictive score of how likely a user can be a troll.

Chapter 8

Future Work

In future, to improve the classification accuracy of the system, as a primary extension, we intend to combine the tweets of an individual user instead of reasoning each tweet separately, to then apply a LDA model and extract a set of topics. In order to model the user behavior, our current system classifies individual tweet and then averages the score of each tweet. The technique lacks the ability to reason the user tweet as a set of topics and uncover the common topics that are discussed over time by the user. By applying LDA model on the collection of tweets, we can do a detailed analysis of the user behavior, reasoning the tweets of the user as a document, than a average similarity of individual tweets. Also the use of LDA allows us to only reason over the set of words that are used, but does not model the conditional dependence between words. Modeling conditional dependence between words used will be crucial for a robust inference, so in the future, we also plan to use n-grams along with LDA to address this issue.

Though the nature of the problem and the lack of required supervision in the dataset heavily accounted for the choice of the unsupervised generative model, we still have not ruled other learning models out in the future. With more data available and better data pruning techniques, we will be able to reliably train discriminative classifiers to perform as well or even better. But as immediate extensions, we intend to compare our methods with other generative techniques such as Naive Bayes. Though Naive Bayes relies on a less flexible distributional model than LDA, it will serve as a concrete baseline to evaluate our model.

We also intend to extend our evaluation section to include more metrics for better representation of the learned model. While the relative similarity score captures the relevance that a tweet or a user account holds over each model, it still does not completely justify how well the model is learned. Use of metrics such as precision-recall or the F-score will allow us to obtain better insight into the method learned.

Bibliography

Russian trolls flood Twitter after Parkland shooting. https://www.nbcnews.com/tech/social-media/russian-trolls-flood-twitter-after-parkland-shooting-n848471. [Accessed: 09-November-2017].

RTwitter deleted 200,000 Russian troll tweets. Read them here. https://www.nbcnews.com/tech/social-media/now-available-more-200-000-deleted-russian-troll-tweets-n844731.

Word2vec - code archive. https://code.google.com/archive/p/word2vec/.

Washington post: The chinese government fakes nearly 450 million social media comments a year. this is why. https://www.washingtonpost.com/news/monkey-cage/wp/2016/05/19/the-chinese-government-fakes-nearly-\450-million-social-media-comments-a-year-this-is-why/?utm_term=.b2642149b152,.

Bristow, michael (16 december 2008) china's internet 'spin doctors. http://news.bbc.co.uk/2/hi/asia-pacific/7783640.stm.

Gallagher, sean (16 december 2008) red astroturf: Chinese government makes millions of fake social media posts. https://arstechnica.com/information-technology/2016/06/red-astroturf-\chinese-government-makes-millions-of-fake-social-media-posts/. Ars Technica Retrieved 2016-06-14.

Cook, sarah; shum, maggie (11 october 2011). china's growing army of paid internet commentators. https://web.archive.org/web/20111013195601/http://blog.freedomhouse.org/weblog/2011/10/chinas-growing-army-of-paid-internet-commentators.html. Freedom House. Archived from the original on 13 October 2011.

Chinese sources used for data collection on 50-cent army. https://www.weibo.com/us,http://dailynews.sina.com/bg/,http://www.qq.com/.

H Allcott and M Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36, 2017.

A Badawy, E Ferrara, and K Lerman. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. *arXiv*, 1802.04291, 2018.

DM Blei, AY Ng, and MI Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, pages 993–1022, 2003.

C Cortes and V Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

A Fourney, MZ Racz, G Ranade, M Mobius, and E Horvitz. Geographic and temporal trends in fake news consumption during the 2016 us presidential election. *CIKM*, 2017.

P Galan-Garcia, JGDL Puerta, CL Gomez, and I P G Santos. Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. *Logic Journal of the IGPL*, 24(1):42–53, 2016.

PN Howard, B Kollanyi, S Bradshaw, and L M Neudert. Social media, news and political information during the us election: Was polarizing content concentrated in swing states? *arXiv*, 1802.03573, 2018.

G King, J Pan, and ME Roberts. How the chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review*, 111(3):484–501, 2017.

T Mihaylov and P Nakov. Hunting for troll comments in news community forums. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2:399–405, 2016.

T Mihaylov, I Koychev, G Georgiev, and P Nakov. Exposing paid opinion manipulation trolls. *In Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 443–450, 2015.

T Mikolov, K Chen, G Corrado, and J Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

C Shao, G L Ciampaglia, O Varol, A Flammini, and F Menczer. The spread of fake news by social bots. *arXiv*, 1707.07592, 2017.

S Zannettou, T Caulfield, E De Cristofaro, M Sirivianos, and G Stringhini. Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web. *arXiv*, 1801.09288, 2018.