Methods of Interference in National Sovereignties and Global Power Distribution:

Analysis of Data Weaponization as means of Hybrid Cyberwarfare

A thesis submitted in partial fulfillment

for the

Bachelor's of Science Degree in Computer Science.

Supervising Professor : Dr. Aristides Vrahatis

Abstract

Social media, such as Twitter, has been taken advantage of and exploited by state actors to manipulate political discourse and spread disinformation during the Clinton v Trump 2016 US Presidential Election. Trolls are users of social media accounts created with the goal to influence the public opinion by posting or reposting messages that contain misleading or inflammatory information with malicious intentions. In this paper, we aim to provide information regarding the characteristics of Russian Troll accounts that have been administered by the Internet Research Agency (IRA), a troll factory allegedly financed by the Russian government, and encourage a form of Russian troll modeling understanding using visualizations, such as but not limited to, histograms, scatter plots for two or more variables, bar plots and grouped bar plots. After modeling the Russian Trolls, we conduct a content analysis of troll tweets to grasp the context of the Russian Troll tweets. Applying several NLP techniques we correlated the Russian troll activity with the understanding of the 2016 US election candidates dynamic and the substance of Russian Trolls' influence on the election. We conducted time and 2-gram-sentiment analysis and presented the observed sentiment polarization and the word that follows each candidate in the dataset's tweets. By using sentiment analysis we showed a measurable difference between the emotions that the 2 presidential candidates were invoking to the trolls which seems to have been reflected in the election results. To visualize the analysis results, we used among others Tornado Charts, countplot charts, Bar charts, Modeling Volatility (GARCH) bar plots, grouped bar plots, word clouds.

Acknowledgements

First and foremost, I would like to thank my mom and dad, who raised me and provided for me without any discounts or conditions and did the best they could so I would be able to have access to the means needed to become the best version of myself.

I would also like to thank my professor and academic supervisor, Dr. Vrahatis, who believed in me, trusted me and gave me **unlimited freedom and support regarding this paper’s topic**, which for our university’s status quo was considered a rather radical and atypical one.

A sincere thank you goes out to my friends who supported my efforts and were there for me in every step of this process, helping me in any way possible.

Last but certainly not least, this thesis paper is dedicated to Hillary Rodham Clinton, whose life journey has inspired millions of people and her **illegitimate and foreign-state assisted** loss during the 2016 US Presidential Elections, contributed into shedding light in a new field of hybrid cyber & information warfare, that revolutionized the way people treat data and social media, in the context of weaponization and opinion shaping dynamics.

Contents

List of Figures

Clinton ‘unfit’ for President of the United States

Abbreviations

Acronym	**What (it) Stands For**
API	Application Programming Interface
US	United States
IRA	Internet Research Agency
AI	Artificial Intelligence
USHPSCI	United States House Permanent Select Committee on Intelligence
NLP	Natural language processing
NER	Named Entity Recognition
ASR	Automatic Speech Recognition
OCR	Optical Character Recognition

Chapter 1: Introduction

The purpose of the introductory Chapter 1 is to familiarize the readers with the issue of spam accounts in Social Media Networks and especially their influence on public opinion and consequently global power dynamic, which was the motivation of this thesis. The structural distribution of the introductory chapter includes a thorough explanation of social media networks, spam accounts, their behavioral patterns on twitter and a description of the research approach that is going to be followed. The chapter concludes with the contribution and novelty of this piece of work.

1.1 Social Media Networks & Spam Behavior : Background

1.1.1 Social Media Networks

In recent years, a rapid proliferation and widespread adoption of a new class of information technologies, commonly known as social media, has been witnessed. People began viewing and treating social media, not just as a way of spending their time but as means of socializing, networking, communicating, sharing important life moments and stating their opinions on crucial issues. Apart from the undeniable impact that social media have had in forging a new way of socialization, they also have created a new market with a current estimated worth of 223.11 $ billion, that is expected to reach a growth of $833.50 billion by 2026 at a compound annual growth rate (CAGR) of 39%[1]. Such social media platforms are free to join and easy to use, easily accessible through their websites or mobile/tablet applications. Despite being theoretically free, social media companies profit by capitalizing on a much more valuable currency, even more valuable than money itself, data, *our data*. Data we create by searching, posting our thoughts, desires and fears and that its worth (big data analytics market) is projected to reach 655.53 billion US dollars in 2029, at a CAGR of 13.4%[2]. Social Networks vary in format and features, from Facebook and Instagram to Twitter and Tik Tok, all have a plethora of different characteristics that

[1] Thebusinessresearchcompany.com. 2022. *Social Media Market Size 2022 And Growth Analysis*. [online] Available at: <https://www.thebusinessresearchcompany.com/report/social-media-global-market-report> [Accessed 25 September 2022].

[2] Insights, F., 2022. *With 13.4% CAGR, Big Data Analytics Market Size Worth USD 655.53 Billion by 2029*. [online] GlobeNewswire News Room. Available at: <https://www.globenewswire.com/en/news-release/2022/07/21/2483358/0/en/With-13-4-CAGR-Big-Data-Analytics-Market-Size-Worth-USD-655-53-Billion-by-2029.html> [Accessed 25 September 2022].

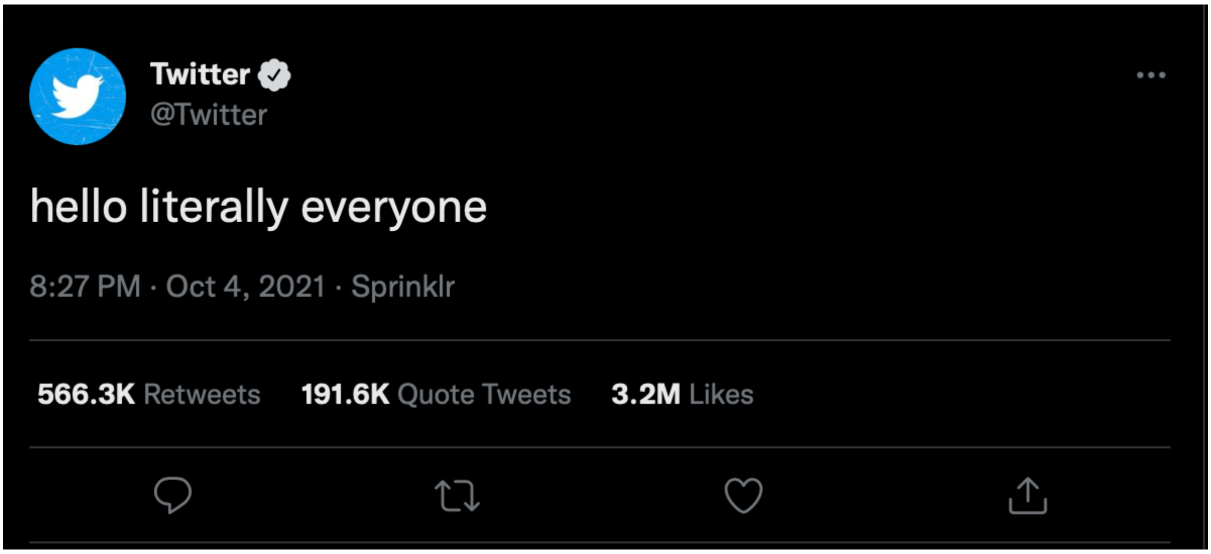
make them alluring to the user. One of the most popular platforms with millions of active users, is *Twitter*.

1.1.2 Twitter: Structure & Features

Twitter is a free social networking tool that provides people with the opportunity to share information in a real-time news feed, by posting comments about their experiences and thoughts. On twitter, users can create their profile, a personalized feed and post "tweets" - texts up to 280 characters long - that can contain images, videos, GIFS, links or any other forms of media. As long as the tweet follows Twitter's policy, there is no restriction regarding the subject one can tweet about. Twitter users, establish online interactions, form connections with other users by following or getting followed by other profiles. The accounts that follow a particular account, are called followers and the accounts that are being followed by a particular account are the followers - or as usually encountered "followings". By following an account the follower "subscribes" to view the content that this profile posts, on demand. It is of uttermost importance to mention that on twitter, contrary to other social media platforms like Facebook, the following functionality is not bidirectional, so the concept of twitter "friends" does not exist officially. However, if two users follow each other, we abusively use the term "friends" to refer to *connected accounts.*

Except for the concept of tweet, Twitter gives the user access to a variety of features. Firstly, twitter allows users to like or *favorite* a tweet, indicating their love or appreciation for it. Twitter, also provides the opportunity to repost a tweet posted and owned by another account, which is called a *retweet.* Accounts can also reply or mention other accounts by adding the "@username" in the text of their tweets. Those actions are known as, *replies and mentions,* respectively. It is important to mention, that only users that have created an account can post, like, reply, retweet other tweets, contrary to unregistered users that can only read publicly available tweets. A twitter user can also utilize something called *hashtag,* which is a word or a phrase that follows the symbol "#" and usually matches the context of the tweet or can be used to make a remark about the tweet itself. To put this into perspective, if someone wants to tweet about a football match, i.e Barcelona F.C VS Real Madrid F.C, or about the results of the US elections, their tweet regarding these topics can contain the hashtags "#RealVSBarcelona" and "#USelections", respectively. By using a hashtag, the tweet that includes the specific hashtag is shown in a forum that contains all the tweets that include this particular hashtag, grouping posts by topic or by type.

Another characteristic that twitter initiated,and then spread to multiple other social media platforms, is the “verified accounts program”. Twitter stated that “the blue badge indicates that an account of public interest is authentic. To receive the blue badge, your account must be **authentic**, **notable**, and **active**”[3] In 2021, twitter stated that the existence of a Wikipedia page will be the criterion and means of verification for accounts to receive the blue badge.



Twitter ✓
@Twitter

hello literally everyone

8:27 PM · Oct 4, 2021 · Sprinklr

566.3K Retweets 191.6K Quote Tweets 3.2M Likes

Figure 1 Example of a tweet

1.1.3 Spam and Spam 2.0

One of the most common issues that individuals have had to deal with since the early days of the internet is *spam.* According to the Anti-Spam Research Lab Digital Ecosystem and Business Intelligence Institute Curtin University, spamming is “the act of spreading unsolicited and unrelated content in several different domains such as email, instant messaging, web pages, Internet Telephony, etc”.[4] Spam, could also be described -in a generic context- as the abuse of electronic messaging systems to send unsolicited messages in bulk indiscriminately[5]. Spam is not something new, as it was first observed in the form of email spam

[3] https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts

[4] Hayati, P., Potdar, V., Talevski, A., Firoozeh, N., Sarenche, S. and Yeganeh, E., 2022. *Definition of spam 2.0: New spamming boom.*

[5] B. Whitworth and E. Whitworth, "Spam and the social-technical gap," *Computer,* vol. 37, pp. 38-45, 2004. Available at: <https://www.semanticscholar.org/paper/Spam-and-the-social-technical-gap-Whitworth-Whitworth/b66b8be6b7438f970a7ac9f25413148101851b8b> [Accessed 25 September 2022].

where unsolicited messages were sent to user's email accounts, since the beginning of the internet era.

Today due to the rapid expansion of the internet and its uses, we have entered an advanced form of Web which is called *Web 2.0*, and is associated with web applications that facilitate interactive information sharing, interoperability, user-centered design and collaboration

on the World Wide Web[6]. It is of crucial importance to underline that Web 2.0's applications are the backbone of the vast majority of web services we use today. On web 2.0, applications' users add value to them, by interacting, participating and by creating a web based community. Functions of Web 2.0 are found in social networking-sites, media-sharing sites, blogs and even folksonomies. They are provided by public, private/personal entities and sometimes by the government itself. This tremendous upgrade of the web and its applications consequently led to the update of what we have defined before as spam, resulting in the genesis of Spam 2.0. "Spam 2.0 is different from the initial spam in the following ways:

- it is targeted at Web 2.0 applications,
- pam 2.0 spreads through legitimate websites such as government, universities, personal homepages etc.
- it can be automatically distributed to as many Web 2.0 sites as possible through the use of automated agents.

The main problems that spam 2.0 can cause, are:

- undeserved high ranking for spammer campaign in search engine results and consequently the low quality content gets higher indexing position than good quality one,
- damage to the reputation of legitimate websites, as it is now deemed as untrustworthy by users and loses its legitimacy,
- waste of valuable resources such as network bandwidth and memory space
- ***and most importantly, tricking users, and damaging the popularity of systems.***"[7]

[6] T. O'Reilly, "What Is Web 2.0," in http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html: O'Reilly Network, 2005. [Accessed 25 September 2022].

[7] P. Hayati, K. Chai, V. Potdar, and A. Talevski, "HoneySpam 2.0: Profiling Web Spambot Behaviour," in *12th International Conference on Principles of Practise in Multi-Agent Systems*, Nagoya, Japan, 2009, pp. 335- 344. [Accessed 25 September 2022].

The rise of Spam 2.0 and its consequences, creates significant socio-political issues, since this not-easily-detectable spam software can be used to influence the public opinion, manipulate the people into making decisions that are not in their advantage and as a result, affect the global power dynamic.

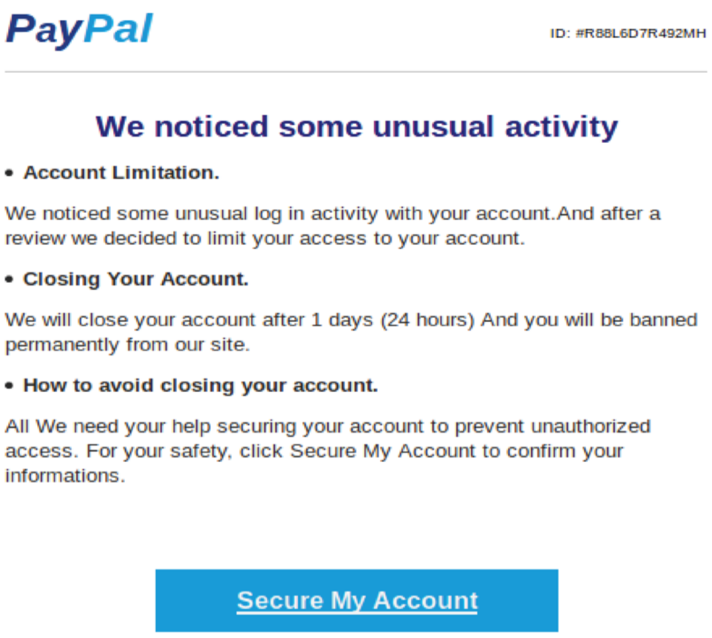
1.1.4 Spam Behavior on Twitter

Following the extreme growth of web applications, social media experienced a germination of equal - if not larger - extent. As a result, *spammers*, who are the users that initiate and generate spam, quickly found their way to social media to expand their spamming behavior. Despite the fact that spamming patterns can be detected in the majority of social media platforms in a plethora of forms and variations, the ones that are found on twitter are the ones that we highlight as of momentous importance. Spam behavior on twitter is easier to occur, since the registration process for someone to join Twitter is easy and does not require the filling of significant input fields by the user. As such, a user can create multiple Twitter accounts by providing different emails & user names. Spammers take advantage of Twitter's easy registration and create a number of accounts in contemplation of expanding their network of influence, increasing the possibility of being noticed by other users on the platform and making the twitter algorithm work in their favor.

Even though we tend to colligate all actors of spam activities under the generic term "spammers", inside twitter there is a miscellany of different types of them. Some are trying to promote themselves, a product or a service to achieve maximum profit, others aim to spread fake news to influence public opinion, others just send follow requests to other twitter accounts to increase the number of their followers and others include URLs inside their tweets, to transfer users to their websites and acquire more audience. Some of those URLs link to websites that may be proven harmful not only for the users but for their computer's hardware as well. Thankfully, ways to identify those malicious URLs, without having to click on them, have been discovered through the years, the most common one being carefully monitoring the structure of the URL strings. A very common example being, replacing a lowercase character with the same capital one, which would lead the user clicking on the URL into a completely different website than they expected.

1.1.5 Spam Behavior & Bots on Web 1.0 as means of financial fraud

In order to maximize the efficacy of their malicious actions, spammers tend to imitate legitimate companies, that type of attack is called *phishing*, because the tweets or emails are used as bait[8]. Paypal's and DHL's customers have been some of the many companies whose brand names were weaponized by spammers. The usual pattern of the emails or other distributed outlets, included a "Dear Customer" as an opening which created a false sense of urgency and was followed by the false statement that there was a high chance that their account was hacked or restricted. To deal with this issue and save their accounts, the spammers' email, urged them to press on the link provided in the end. In the DHL case, which is a shipping-delivery company, the customer is contacted via message (email or SMS) and is told that in order for their package to be delivered they need to pay an extra amount of money by filling out their credit card info, in a webpage format looking exactly like the official one. Both Paypal and DHL have put out thorough instructions and warnings on their official sites to educate users about the characteristics of the spam mails and protect them from phishing and similar spamming frauds.[9]



PayPal ID: #R88L6D7R492MH

We noticed some unusual activity

- **Account Limitation.**

We noticed some unusual log in activity with your account.And after a review we decided to limit your access to your account.

- **Closing Your Account.**

We will close your account after 1 days (24 hours) And you will be banned permanently from our site.

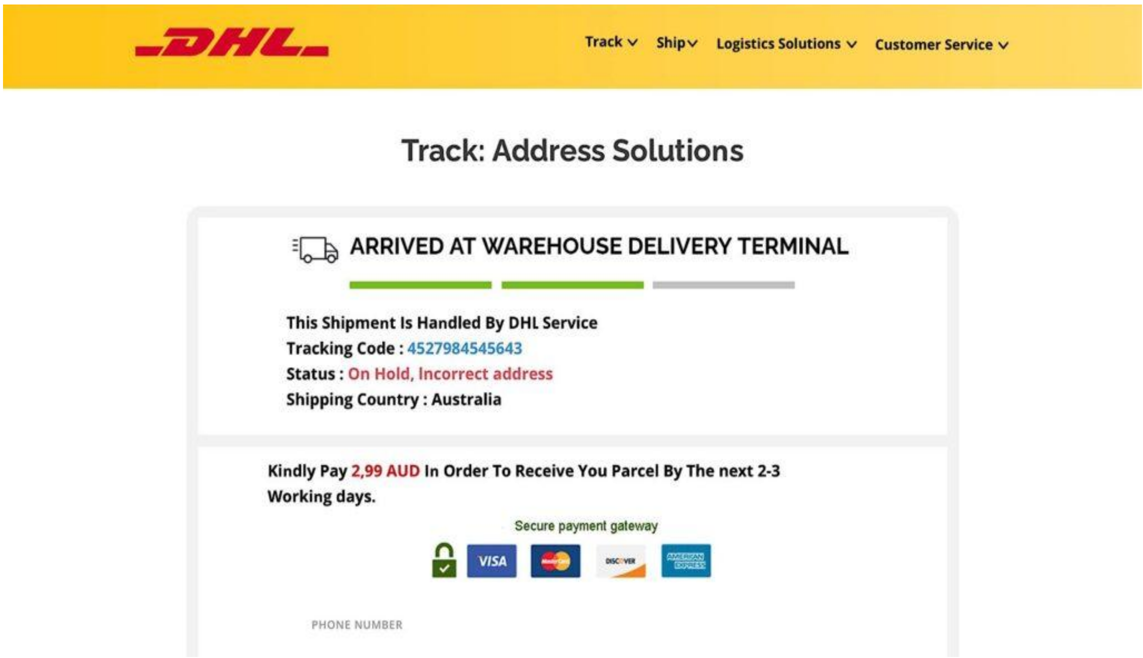
- **How to avoid closing your account.**

All We need your help securing your account to prevent unauthorized access. For your safety, click Secure My Account to confirm your informations.

Figure 2 Example of a spam email on Paypal

[8] "How to Spot a Fake PayPal Email," [Online]. Available: https://www.secureworldexpo.com/industry-news/how-to-spot-a-fake-paypal-email. [Accessed 25 September 2022].

[9] https://www.dhl.com/gr-en/home/footer/fraud-awareness.html [Accessed 25 September 2022].



Track: Address Solutions

ARRIVED AT WAREHOUSE DELIVERY TERMINAL

This Shipment Is Handled By DHL Service
Tracking Code : 4527984545643
Status : On Hold, Incorrect address
Shipping Country : Australia

Kindly Pay 2,99 AUD In Order To Receive You Parcel By The next 2-3 Working days.

Secure payment gateway

PHONE NUMBER

Figure 3 Example of a spam message pretending to be sent by DHL

Over time,users began to realize the spamming patterns that could result in fraud and got more vigilant. However, when spammers realized their actions were no longer effective, also developed mechanisms to overcome such obstacles. This perpetual pursue of spammers by users, social media platforms and the authorities has undoubtedly proclaimed

1.1.6 Twitter Bots: Behavior & Distribution

Another essential feature of Twitter is that it provides users with a number of automation tools allowing them to post tweets by providing the context of the tweet beforehand, through programming scripts that can be customized to fill the needs of the user. Twitter also provides the auto-searching and auto-following users functionality. These options have resulted in the creation of another kind of users which are called *bot users or bots.* Bots are being created and used by spammers to maximize their success rate and the quantity of content posted per minute. That is achieved by 1) taking advantage of the auto-posting feature to schedule their tweets and 2) by randomly searching and following other users to eventually expand their network.

A Bot is a type of user that is not in actuality a person but a machine that operates mechanically to emulate user interaction of twitter users.[10] The issue with bots arises when they begin to imitate people[11] in order to deceive or manipulate other users.

The most crucial case of this kind is the derailment of political conversation back in the United States 2016 elections where misinformation and divisive messaging played a major role in the outcome[12] of this election, something that we will thoroughly investigate and prove as we move forward. Other cases include, the proliferation of hate speech and racist rhetoric propagated through them or even with misinformation in the recent COVID-19 epidemic[13]. These bot users can be created for a variety of goals, purposes that determine the kind of bot they are.

1) Cyborg: Accounts that have both bot and human activity with one aiding the other.
2) Spambot: As the name implies, bots designed to produce and spread spam to users.
3) Social bot: Bots that operate in the effort to influence the course of discussion on social media and attract followers.
4) **Political bot: Social bots used for political campaign efforts.**
5) Other bots: Other types of simple bots.

The main media conglomerates such as Facebook, Google and Twitter have joined the fight against bots and their illegal activities. Their main strategy against these entities is removing the detected fake accounts.For instance, Twitter and Facebook have blocked a campaign supported by the Chinese authorities conducting propaganda operations concerning the protests in Hong Kong in 2019. Twitter deleted 2,000,000 fake accounts that were marked as trolls, in 2018 alone.

1.1.7 Trolls: Definition, Tactics & Differences with Bots

In the era of Internet and social media, there are about 3.8 billion active social media users and 4.5 billion people accessing the internet daily. Each year there is a nine percent

[10] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. (2016). "The rise of social bots." Commun. ACM 59, 7 (July 2016), 96–104 [Accessed 25 September 2022].

[11] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Alessandro Flammini, and Filippo Menczer (2017) "The spread of fake news by social bots"[Accessed 25 September 2022].

[12] Bessi, Alessandro and Ferrara, Emilio (2016), "Social Bots Distort the 2016 US Presidential Election Online Discussion" First Monday, Volume 21, Number 11 -- 7 November 2016 [Accessed 25 September 2022].

[13] Himelein-Wachowiak, M., Giorgi, S., Devoto, A., Rahman, M., Ungar, L., Schwartz, H. A., Epstein, D. H., Leggio, L., & Curtis, B. (2021), "Bots and Misinformation Spread on Social Media: Implications for COVID-19. Journal of medical Internet research" [Accessed 25 September 2022].

growth in the number of users and half of the internet traffic consists of mostly bots and trolls.

Having defined and elaborated upon bots, it's of crucial importance to insert the term *troll* in our analysis. A troll has been defined in various ways, but it is most predominantly defined as a person or a computing system that initially pretends to be a legitimate participant in an online discussion, but later on tries to disrupt the communities coherence[14] - "someone or something who intentionally disrupts online communities"[15]. In a broader way, we can define trolling as "negatively marked online behavior"[16]. In this paper, we adopt a definition of trolling that includes the use of social media - and specifically twitter - to spread false information about social issues and people of power, to affect the public opinion.

It is argued by a plethora of scholars, that trolling is an umbrella term for a spectrum of multifaceted, antagonistic or deviant behaviors online[17]. However, there is pointed out especially by feminist and anti racist scholars that trolling can often be a form of identity-based harrasement especially in the context of invalidating someone's opinion on Twitter discussions, purely by their gender or race.[18]Literature demonstrates that trolling can be a range of behaviors from hacking, to releasing private information, posting satirical comments, posting redundant information in order to disrupt a conversation,or to hate speech.

Trolls are considered to be highly inflammatory accounts that actively work to anger and violate others. Despite the fact that this is often the case, not all trolls operate that way.

The most malicious trolls do not immediately expose their true intentions or characteristics, but work slowly to gain support from individuals from all over the ideological spectrum and push them further into their already solidified beliefs. In this way, they do not initiate fights *but they slowly maintain online*

[14]6. Donath Judith S. Identity and deception in the virtual community. *Communities in Cyberspace* 1999 [Accessed 25 September 2022].

[15] Schwartz Mattathias. *NY Times Magazine.* 2008. The trolls among us.

[16] Hardaker Claire. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *J Politeness Res* 2010 [Accessed 25 September 2022].

[17]Gray, K. L. (2017). Gaming out online: Black lesbian identity development and community building in Xbox Live. Journal of Lesbian Studies, 22(3), 282–296. https://doi.org/10.1080/10894160.2018.1384293 [Accessed 25 September 2022].

[18] Gray, K. L., Buyukozturk, B., Hill, Z. G. (2017). Blurring the boundaries: Using Gamergate to examine "real" and symbolic violence against women in contemporary gaming culture. Sociology Compass, 11(3), Article e12458. [Accessed 25 September 2022].

polarization. Initially, the act in a friendly approachable way to insert themselves into the online community - especially twitter - and gain a following to influence later. Similar to bots, trolls, actively spread disinformation online and dichotomize audiences.

Bots and Trolls are similar in many aspects, but key differences between the two categories exist. Analyzing and presenting them, will act as a main component in reducing their negative impact on online communities and consequently, on people that may fall for their malicious patterns. The main difference between bots & trolls, is that bots - contrary to troll accounts that are fake accounts controlled by humans - are automated social media accounts programmed to perform actions to mimic humans (e.g liking, sharing, or commenting on posts). Furthermore, a bot, as was mentioned before,does not always have malicious purposes, since there are bots that can be helpful or entertaining. For example, some are programmed to help Twitter users condense threads into clickable links (@theadreaderapp) or share pictures of Earth from space (@dscovr_epic). On the other hand, trolls' sole purpose is to disturb online communities by posting inflammatory messages or off-topic comments -although, the most sophisticated trolls attempt to be friendly and not aggressive online. Both bots and trolls can work either independently or coordinate as an extensive network, which is called botnet if it is in regards with bots.[19]

Both trolls & bots have a tremendous impact on affecting public opinion in a superfluity of issues. In 2020 researchers from Carnegie Mellon University[20], reviewed 200 million tweets about COVID-19 - measures, stay-at-home orders, reopening of the US services etc - and they figured that *82% of the top most influential retweeters were bots, 62% of the top 100 retweeters were also bots and that bots fueled 50% of the entire conversation.* The high number of fake accounts that exist online has a direct effect on the understanding of online conversation about people's brands, likeability or acceptance ratings. Both of them can be enlisted to spread false information about issues, constitutions or people, intentionally deceive audiences into believing false information by providing them with fake evidence, increase difficulty of providing accurate information to audiences and downsizing the credibility of the network that is posted on.

1.1.8 Troll Factories

[19] https://link.springer.com/chapter/10.1007/978-0-387-44599-1_8

[20] https://www.cmu.edu/news/stories/archives/2020/may/twitter-bot-campaign.html [Accessed 25 September 2022].

In Internet communication, a troll is defined as a person who provokes disputes, e.g. by raising controversial topics or attacking other participants. However, a troll factory is an entity conducting disinformation propaganda activities on the Internet. This activity is often concealed under an inconspicuous name, e.g. public relations agency, *Internet Research Center*, etc. Troll factories usually operate on the political or economic sphere. The purpose of these operations may be attacking political opponents, unfairly attacking a competing company or other action indicated by the ordering party. Troll factories fulfill their purposes by using and weaponizing, among other things, fake news and hate speech.[21]

"Troll factory", as a term, began to come up on a broader scale in media reports in 2015.[22] At that time, the existence of a troll factory in St. Petersburg employing 300 people was revealed. The agency operated as "the Internet Research Agency", managed by a Russian oligarch, Yevgeny Prigozhin. Employees' duties included publishing on the Internet, mainly on social media, posts praising Russian President Vladimir Putin, and criticizing countries not supporting Russia.

Employees of troll factories are in charge of creating fake identities and running those profiles on social media[23]. Their main goal is to give the illusion of authenticity of the created account, therefore, these accounts do not only post the content related to the purpose of the troll factory but also material that gives a sense of credibility to the fake account, such as posts concerning private life issues. The profiles feature images taken from ready-made Internet repositories, altered to mislead Internet search engines. Since the troll factory employs hundreds of people and each employee has several accounts, it is easy to create a social network linking fake profiles, interacting with each other and creating the impression of a real community. The longer the given accounts are maintained, the more real they tend to appear, as more time is given to formulate their fake backgrounds. Troll factories' employees work in shifts to make sure that the messages they produce are displayed on a 24-hour basis.

While troll factories' most well known activity is creating fake social media profiles, they create other products as well to maximize credibility to their posts or accounts. Troll factories

[21] Aro, J. (2016), The Cyberspace War: Propaganda and Trolling as Warfare. Tools, European View, (15), 121–132. [Accessed 25 September 2022].

[22] Duskaeva, L.R., Konyaeva, L.R. (2016), Trolling in Russian Media, Journal of Organizational Culture, Communications and Conflict, (4), 58–67. [Accessed 25 September 2022].

[23] Karpan, A. (2018),Troll Factories: Russia's Web Brigades, Greenhaven Publishing, New York. [Accessed 25 September 2022].

can go as far as creating entire websites and blogs to support their trolling operations and give the illusion of legitimacy to their posts. Except for creating posts, troll factories' employees also respond to comments to take part in online discussions. They can also simulate disputes in order to increase the impression of authenticity of their fake profiles[24].

Troll factories are actors that operate in a planned manner, always in accordance with the instructions and recommendations of the ordering party, who pays for their coordinated posting. Internet trolls' activities are paid for and controlled by top-down guidelines. After the content is created by them, it reaches "volunteers" who continue their propaganda and disinformation operations on their own free will, without any compensation, because the messages distributed by troll factories are in line with their political views or in general supported ideology.[25]

It is of utmost importance to underline that internet trolls' operations are supported by bots, so there is indeed a direct interconnection between their activities. Those activities can vary from programs sending out messages automatically to automated responses to the appearance of certain keywords. While the messages sent by bots demonstrate a low degree of reliability and are easy to be detected and classified as spam, due to language errors, duplication of messages or other mistakes, the material distributed by troll factories may appear more reliable, and therefore their disinformation activities will be more effective.

1.2 Topic of Bachelor's Thesis

1.2.1 State of the art - Topic Statement

It is a common belief in the scientific community that Russia's Internet Research Agency (IRA) attempted to interfere with the 2016 U.S. election as well as other elections by running fake accounts on Twitter - commonly known as "Russian troll" accounts. This intervention could have vast consequences considering the viral nature of some tweets, the quantity of users exposed to Russian trolls' content and the substantial role social media have played in political campaigns of the past.

[24] Bernal, P. (2018), The Internet, Warts and All: Free Speech, Privacy and Truth, Cambridge University Press, Cambridge. [Accessed 25 September 2022].

[25] Lehto, M., Neittaanmäki, P. (Eds.) (2018), Cyber Security: Power and Technology, Springer, Cham. [Accessed 25 September 2022].

In May 2018, the Democratic representatives from the United States House Permanent Select Committee on Intelligence (USHPSCI) published their findings regarding Russian interference in the 2016 United States presidential election. In their report, the Committee supported and reaffirmed conclusions drawn by the Intelligence Community regarding election interference taken by the Kremlin, ranging in scope from hacking -and- dumping campaigns to the dissemination of propaganda. Additionally, the Committee's report revealed several details resulting from further investigation by the Internet Research Agency (IRA), a Saint Petersburg-based company known to have engaged in long-term influence operations on behalf of Russian political and economic interests.

According to data provided to US Senate's Intelligence Committee[26] by Twitter, a snapshot of relevant Twitter activity in the period between September 1 and November 15, 2016 reveals:

- More than 36,000 Russian-linked bot accounts tweeted about the U.S. election
- More than 36,000 Russian-linked bot accounts tweeted about the U.S. election
- Approximately 288 million impressions of Russian bot tweets; and
- More than 130,00 tweets by accounts linked to the IRA.

According to the February 2018 US grand-jury indictment the Internet Research Agency (IRA) in St. Petersburg, Russia administered "information warfare" against the United State by using fictitious U.S. personas on social media platforms.." (US Department of Justice, 2018). These fake personas communicated with unsuspected members of the public **to inspire distrust in the political system, discourage minorities from voting, make assertions of voter fraud, organize political rallies, stoke racial divisions** (Senate Intelligence Committee, 2019), and other illegal activities (US Department of Justice, 2018).

As per the Senate Intelligence Committee's **bipartisan report**, the Russian Federation interfered to provide assistance to the Trump campaign, and eventually to help him win the 2016 US presidential elections:

"The Committee found that the IRA sought to influence the 2016 U.S. presidential election by harming Hillary Clinton's chances of success and supporting Donald Trump at the direction of the Kremlin. The Committee found that IRA social media activity was

[26] https://intelligence.house.gov/social-media-content/ [Accessed 25 September 2022].

overtly and almost invariably supportive of then-candidate Trump to the detriment of Secretary Clinton's campaign."

A consequence of the USHPSCI report has been the public release of two datasets reflecting the behaviors of IRA actors at the time of this writing. The second dataset, which contains timelines for over 1,200 English-language Twitter accounts that were found to be operating as agents of the IRA, was curated and released in July 2018 by Darren Linvill and Patrick Warren The data set contains fruitful information in the terms of timelines, behaviors, and the language used by Russian Troll accounts and associated metadata.

According to the IRA US Department of Justice, Internet Research Agency employees used Internet proxy services to conceal their I.P. addresses and not be able to be traced and located back to the Russian Federation. The stealth of this operation has made it almost impossible to know what was the impact of it - if there was any, to begin with. This point becomes clear, especially in the "downstream" consequences, which were the specific path of content retweeting, sharing or replying etc., and their behavioral effects and decision influencing properties cannot be traced explicitly.

Empirically, the activities of the professional Russian Trolls in question, are in rigid correlation with the timing of overall internet traffic levels and legitimate events and especially breaking news stories that can attract much desired traffic. A prime example of the link between events and trolls tweeting is the "the well-known faint/stumble by Hillary Clinton leaving a 9/11 commemoration event, followed by her pausing the campaign with an announcement of pneumonia". It is of crucial importance to point out that trolling activity follows time trends, seasonality, and the influence of the day of the week. That specific characteristic makes it harder to attribute time-series disparities in trolling magnitude and its effects to the work of trolls, in contrast to other factors.

In this paper, we present an initial analysis of the published IRA data - focusing on Twitter timelines that have been publicly released. Our main goal is to characterize the social media behaviors of the IRA prior to, during, and following the 2016 election. It is our hope that these findings will serve to facilitate continued and deeper investigations of foreign interference in both past and future democratic processes and will provide further proof that **data can be used as means of destabilizing democracies and affecting global power distribution.**

By conducting a modeling analysis of Russian trolls' tweets, detecting and visualizing common patterns in their behavioral schemes and demonstrating a number of characteristics they share, to connecting russian troll tweets with breaking news in the political scene, we will attempt to assist the research community and the general public in understanding the foundational details of when and how the IRA attempted to manipulate the psychological landscape surrounding the election.

1.2.2 Twitter & Platform Manipulation

Twitter provides a plethora of guidelines regarding spam and platform manipulation[27]. According to Twitter, Platform Manipulation includes:

- "commercially motivated spam, that typically aims to drive traffic or attention from a conversation on Twitter to accounts, websites, products, services, or initiatives
- inauthentic engagements, that attempt to make accounts or content appear more popular or active than they are
- coordinated activity, that attempts to artificially influence conversations through the use of multiple accounts, fake accounts, automation and/or scripting."

1.2.3 Penalties for Spam Behavior

Twitter aims to create a platform that can be seen as a place where people can have human connections, find reliable information and express themselves freely and safely, as stated in their Rules and Policies. To achieve that Twitter relies on custom-built tools to identify and deal with that behavior.

In that context, they have improved the phone verification feature and introduced new safety features such as reCAPTCHAs, to help ensure that a human is in control of an account. The consequences of violating policies depend on the severity and the kind of violation, as well as any previous history of violations. To protect the users twitter has proceeded to a series of actions:

- **Anti-spam challenges**

[27] "Platform manipulation and spam policy," https://help.twitter.com/en/rules-and-policies/platform-manipulation [Accessed 25 September 2022].

When Twitter detects suspicious levels of activity, accounts may be locked and prompted to provide additional information (e.g., a phone number) or to solve a reCAPTCHA.

- **Blacklisting URLs**

Twitter provides warnings and blacklists URLs that are considered to be unsafe and may lead the user that presses on them on sites that can violate their rights.

- **Tweet deletion and temporary account locks**

In the case that the platform manipulation or spam offense is an isolated incident or first offense, the actions range from requiring deletion of one or more Tweets to temporarily locking the account or accounts that violated the rules. Any further platform manipulation offenses will result in permanent suspension.

In the case of a violation centering around the use of multiple accounts, the user might be requested to preserve only one account. The remaining accounts will be permanently suspended.

- **Permanent suspension**

For severe violations, accounts will be permanently suspended at first detection. As severe violations Twitter considers severe violations operating accounts that the main behavior and tactics are in violation of the aforementioned policies, utilizing any means and **tactics to undermine the integrity of elections**, buying or selling accounts and operating accounts that are attributed to entities that are known to violate Twitter’s rules.

1.2.4 European Union General Data Protection Regulation

The regulation in Article 22 about “Automated individual decision-making, including profiling” states:[28]

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

2. Paragraph 1 shall not apply if the decision:

[28] https://ec.europa.eu/info/law/law-topic/data-protection/eu-data-protection-rules_en#gdpr-the-fabric-of-a-success-story

(a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;

(b) is authorized by Union or Member State law to which the controller is subject, and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or

(c) is based on the data subject's explicit consent.

3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9 (1), unless point (a) or (g) of Article 9 (2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

According to Article 4 profiling is "Any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyze or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behavior, location or movements". Based on Article 22, it is made clear that the GDPR restricts anyone from making solely automated decisions, including those based on profiling, that have a legal or similarly significant effect on individuals. For something to be solely automated there must be no human involvement in the decision-making process.

1.3 Thesis Contribution & Challenges

1.3.1 Selection of Twitter Social Media Platform

Twitter, alongside Facebook and Instagram, is considered to be on the top three most popular social media platforms worldwide, with more than 330 million active users. Due to its broad use and the automation that Twitter's API provides, spam and bot accounts are a common phenomenon in comparison with the aforementioned platforms. Another factor that contributed to the selection of Twitter as the social media platform of this paper, is the plethora and variety of research and previous work on the subject that exists in literature and also the number of labeled datasets that are publicly available for use. Furthermore, the fact that Twitter is one of the few social media platforms that provides an API cannot be hushed up as a crucial contributor in the selection of Twitter as the platform on which we conduct our analysis.

1.3.2 Machine Learning: Text Mining

Text Mining has developed over decades and across academic sectors to create a diverse body of literature on the computed-aided analysis of data derived from texts (textual data). Fields that can text mining can be implemented in fields such as but not limited to, statistics, political science, computer science linguistics and computer science in general[29]. As a result of the multidimensionality of text mining, different terminology has evolved including computational content analysis and natural data language processing. In spite of their different emphasis, they share an obvious focus on computed supported processing and analysis of text in the form of natural language, using automated means. Due to this plethora of text mining implementations in many disciplinaries a consensual definition of text mining remains absent. However, a broad agreement on the generic procedure of analyzing big-scale text-data, in terms of knowledge discovery in data sets[30].

According to Miner (et al. 2012), the unifying theme that links the numerous text mining techniques is the concept of converting unstructured data that is in text form, into structured data in the form of numbers, so statistical and mathematical algorithms can be applied to it. A typical approach is using the form of a matrix, to represent a text document. In this matrix, each column constitutes a document and each row a specific term and the cells contain the frequency of the term in the specific document.

That approach faces two major problems. Firstly, one of the essential characteristics and a potential issue of representation of textual data in such a document-term matrix, is that it neglects the language’s structure in the document and treats it as a sum of words. Moreover, vast collections of documents can catalyze the creation of a document-term matrix of a very large number of dimensions. To deal with these limitations, text

[29] Beinke, Thies & Schamann, Annabell & Freitag, Michael & Feldmann, Klaas & Brandt, Matthias. (2017). Text-Mining and Gamification for the Qualification of Service Technicians in the Maintenance Industry of Offshore Wind Energy. International Journal of e-Navigation and Maritime Economy. 6. 44-52. 10.1016/j.enavi.2017.05.006. [Accessed 25 September 2022].

[30] https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf

preprocessing techniques have been developed in various fields such as computer and information science and linguistics[31].

Text Preprocessing

Text preprocessing begins with a procedure called tokenization, which is the process of converting a sequence of characters into tokens. Words, sentences or paragraphs can be used as tokens and seperate strings at white spaces. Using tokens and converting them into a document-term matrix yields the above-mentioned issue of disregarding the linguistic structure and treating the document as a bag of words. However, this approach is too simplistic to take account of the complex linguistic structure of compound words such as "innovation culture" or "open innovation" that used together have a specific different meaning. Algorithms to detect such word combinations are now available, and are known as n-grams. Before forming a document-term matrix, n grams have to be replaced so the essence of the document is not lost. As a result, the first issue is tackled.

The second major problem is the formation of large (high dimensional) document-term matrix results because of the variability of human language. For example, we use words in temporal, plural or otherwise inflected forms, in automated analysis this increases tremendously the size (dimensionality) of the document-term matrix since each variation would create a different column or row to the matrix. A plethora of processing techniques have been developed to reduce such variability while at the same time preserving the world's essence and meaning, such as stemming and lemmatization that reduce words to their stem or their lemma, which is the way that the word is found in a lexicon[32]. Some techniques convert all text to lower cases, remove every punctuation point, eliminate all pronouns and function words that do not carry any meaning[33] (e.g 'the', 'and'), others cross out infrequent words[34] or weigh words using criteria such as

[31] Rüdiger, Matthias & Antons, David & Salge, Oliver. (2017). From Text to Data: On The Role and Effect of Text Pre-Processing in Text Mining Research. Academy of Management Proceedings. 2017. 16353. 10.5465/AMBPP.2017.16353abstract.

[32]https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

[33] Antons, David & Grünwald, Eduard & Cichy, Patrick & Salge, Oliver. (2020). The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. R&D Management. 50. 10.1111/radm.12408.

[34] Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1 - 30.Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1 - 30.

term frequency and document frequency, to spot words that discriminate documents from each other.[35]

Except for dealing with the aforementioned problems, preprocessing can also add information in the form of tags to words or phrases, that might provide help to the programmer to pick certain text elements or to structure texts.

After preprocessing the textual data, we move to the phase of actual computer-aided content analysis utilizing either dictionary based techniques or algorithmic techniques.

Computational content analysis with dictionary-based techniques - Sentiment Analysis

Text Mining techniques that are based on word frequency and rely on word frequency counts to calculate contextual, psychological or semantic concepts and constructs, are among the most widely adopted approaches regarding the computational analysis of textual data in research so far.[36]

Researchers may utilize pre existing dictionaries that were developed in previous researches or create new dictionaries that are in accordance with their projects' needs. According to Gamache et al. 2015, theory can be used to construct definitions and survey items to build a dictionary for the construct of regulatory focus[37]. In an epistemological context, dictionaries for text mining can be formed deductively based on existing theory inductively from the already given core or by combining the two aforementioned techniques. Specifically to measure general positivity, negativity or specific emotions in text, validated dictionaries already exists[38].

Dictionary-based text mining is referred to as sentiment analysis and of the most popular lexicons for sentiment analysis is Afin.

1.3.3 Python

Python is considered to be one of the fastest-growing programming languages that includes not only inbuilt but also third-party libraries and packages, suitable for performing text mining.

[35] Antons, David & Breidbach, Christoph & Joshi, Amol & Salge, Oliver. (2021). Computational Literature Reviews: Method, Algorithms, and Roadmap. Organizational Research Methods. 109442812199123. 10.1177/1094428121991230.

[36] https://academic.csuohio.edu/kneuendorf/c63310/Shortetal10.pdf

[37] Gamache, Daniel & Mcnamara, Gerry & Mannor, Michael & Johnson, Russell. (2014). Motivated to Acquire? The Impact of CEO Regulatory Focus on Firm Acquisitions. The Academy of Management Journal. 58. 10.5465/amj.2013.0377.

[38] Pennebaker, James & Boyd, Ryan & Jordan, Kayla & Blackburn, Kate. (2015). The Development and Psychometric Properties of LIWC2015. 10.15781/T29G6Z.

Regarding its Natural Language Toolkit (NLTK), python includes, among others, a NLTK package that contains numbers, Natural Language Processing methods like word and sentence tokenization and classification. This toolkit is basically a collection of programs and libraries for statistical language processing for python languages. This tool is used by a plethora of companies and organizations for multiple purposes such as artificial intelligence, business analytics, market analysis, Natural Language Processing and software making. Packages like Gensim and Spacy, are used for more versatile advanced text mining applications.

1.3.4 Contribution

Investigations on the part of multiple agencies/agents have come to the overwhelming conclusion that Russian Interference in the 2016 U.S presidential election. As part and consequence of recent reports, multiple datasets that capture the actions of the Internet Research Agency (IRA) employees, have become public. The topic of russian trolls trying to influence and shape the public opinion in according with another state's interests and consequently interfere with the electoral procedure, has been well researched and many papers are being published and will continue to be published in the years to come, as it is considered the first prime example - or patient zero - of an cyberwarfare act that utilized data and opinion mining that achieved influencing the electoral results and consequently, global power distribution. In this paper, we present an analysis of Twitter troll accounts actions and characteristics, attempting to model them and find common traits in their actions and behavioral patterns, through the use of language analysis and Visualization. Addionality, we utilize text mining, a plethora of N-grams - specifically Bi-grams - and other content and sentiment analysis techniques to monitor and characterize the evolution of the IRA content before, over and after the course of election circle and correlate it with crucial political events and reactions of the American Electorate.

1.4 Document Structure

This document has 6 chapters in total, including the introductory chapter, in which the readers are introduced briefly to Social Media Platforms basics, along with definitions about the issue of trolls and bots on social media platforms and a selection of methods on approaching bot/troll detection.

In **Chapter 2**,, an overview of related works is presented.

In **Chapter 3**, the methodology followed in the paper is analyzed and broken down into its core pieces.

In **Chapter 4**, the necessary technological knowledge is provided in order for the readers to understand the development's flow and the choices made during the process of the overall system's implementation. Additionally, the visualization of the results is presented and explained.

In **Chapter 5**, conclusions of this thesis are made, and ideas are suggested for further research and future work.

In **Chapter 6**, the references and bibliography used for this paper are mentioned.

Chapter 2 Related Work

The concept of parthenogenesis is considered rare if not utopian in research. Consequently, there has been previous studies that have focused on russian trolls, their interactions with humans, the classification of them as bots or not and even their interference with elections.

Yet, to our knowledge, no studies have examined both the common characteristics of russian troll accounts and whether these efforts actually impacted the attitudes and behaviors of the American public.

2.1 Russia's Interference on Political Campaigns

While state-level online interference in democratic processes in the context of cyber and information warfare is an emerging phenomenon, research has proven Russian's online manipulation campaigns in countries other than the United States. Earlier research has shown, for example, a high amount of Russian tweets were produced in the week prior to the voting day of the 2016 EU (Brexit) Referendum, and then dropped afterwards (Talavera, Pham, and Gorodnichenko 2018). The MacronLeaks campaign that occurred during the 2017 French Presidential elections and the Catalan Referendum, are both political incidents that the Russian trolls were involved with (Stella, Ferrara, and De Domenico 2018).

2.2 Emerging Work on Russian Trolls

Despite the fact that this particular area is a new sector of scholarship, emerging work has examined the datasets of Russian trolls released by twitter and validated by the Senate's Intelligence Committee. Researchers associated with Clemson University identified 5 categories of trolls and stated that the behavioral patterns between trolls distributed in these categories were radically different.(Boatwright, Linvill, and Warren 2018)[39] They were specifically marked as left or right leaning and the dataset contained both. For example, the IRA promoted more left-leaning content on Facebook[40], while right-leaning Twitter handles received more engagement (Spangher et al. 2018).

Furthermore, new research has examined how Russian Troll accounts' tweets were retweeted in the context of the #BlackLivesMatter[41] movement (Stewart, Arif, and Starbird 2018), that was heavily targeted by the trolls in question. Once again, the retweets were divided into different political perspectives, trying to influence all sides of the ideological spectrum. The division and political polarization that has emerged in the last decade, has been taken advantage of by the trolls, who have actively tried to intensify it.

The academic society has not found consensus on whether or how the Russian Trolls are predictable in their actions. Zannetou et al. 2018b, shows that trolls' tactics and targets change overtime, implying that the task of automatic troll detection is not simple and needs further research and more upgraded detection models, while Griffin and Bickel (2018), argue that Russian trolls are composed of accounts with common but customized behavioral characteristics that can be utilized for future troll identification.

Researchers have also examined actual users who have interacted or interact with Russian Troll accounts on Twitter. For instanced, it is proven that, as far as ideological background is concerned, misinformation produced by the russian trolls on Twitter was shared and retweeted more by people identifying as conservatives, rather than liberals.(Badawy, Ferrara, and Lerman 2018). According to Badawy, Lerman, and Ferrara's 2018 research political ideology, bot likelihood scores and activity-related metadata can be utilized for the formation of models that predict

[39] Darren L. Linvill & Patrick L. Warren (2020) Troll Factories: Manufacturing Specialized Disinformation on Twitter,Political Communication, 37:4, 447-467, DOI: 10.1080/10584609.2020.1718257

[40] https://arxiv.org/abs/1810.10033

[41] Ahmer Arif, Leo Graiden Stewart, and Kate Starbird. 2018. Acting the Part: Examining Information Operations Within #BlackLivesMatter Discourse. Proc. ACM Hum.-Comput. Interact. 2, CSCW, Article 20 (November 2018), 27 pages. https://doi.org/10.1145/3274289

which users will spread the misinformation spread by russian trolls.

Regarding the effect of russian trolls’ propaganda, researchers have also worked to clarify the influence of the Russian trolls’ propaganda attempts on social media platforms by utilizing data from Facebook ads, IRA related tweets on Twitter and log data from browsers. The results clearly indicate that 1 in 40.000 web users were exposed to the IRA ads on any given day, but there was a differentiation between right and left leaning content (Spangher et al. 2018). The influence of the Russian trolls has been measured in platforms like Twitter, Reddit and Gab[42] by utilizing Hawkes Processes, according to Zannetou (Zannettou et al. 2018b).

2.3 Bots

Despite the fact that bots were initially created to assist helpful procedures, such as auto replies, they can also become harmful, as they can be weaponized in order to steal users personal information on social platforms (Ferrara et al. 2016), and also spread propaganda. (Shao et al. 2017; Gorodnichenko, Pham, and Talavera 2018). Previous research has indicated that bots largely intervened with the 2016 US presidential Elections, as bots were behind millions of tweets the week before the 2016 US election dates (Bessi and Ferrara 2016). A major disinformation campaign was also conducted before the 2017 French Election (Ferrara 2017). Current attempts to detect bots on social media include systems based on crowdsourcing and human intelligence, and machine learning methods using indicative features (Ferrara et al. 2016). However, previous findings indicate that is becoming harder to filter out bots due to their sophisticated and adaptable behavior (Subrahmanian et al. 2016)

2.4 Deception and Identity Online

Russian trolls have attempted to mask their identities on twitter always in accordance with the audience they want to influence and the topic they are tweeting about. For instance, bots were detected pretending to be African - American activists supporting the #BlackLivesMatter movement (Arif, Stewart, and Starbird 2018). Seminal research, has shown the significance of the essence of “identity” and its influence, varies in different online communities, as the costliness of faking certain social signals is indissolubly connected with their trustworthiness, an insight that researchers use to compose quantitative features (Donath 2002). The significance and salience of identity signals

[42] https://arxiv.org/abs/1801.09288

and also the possible deception caused by them, can be detected in all social media platforms. Myspace users listed books, movies, and TV shows in profiles to build elaborate taste performances in order to convey prestige, differentiation, or aesthetic preference (Liu 2007). On Twitter[43], users manage their self-representation not only via their personal profiles but also via their interactions and participation in ongoing conversations (Marwick and Boyd 2011).

Chapter 3 Methodology

3.1 Dataset Selection

Context

In October 2018, Twitter released 2.9 million English-language tweets from 3,841 accounts as "affiliated with the IRA".

In its October 2018 announcement, Twitter wrote:

"We are releasing the full, comprehensive archives of the Tweets and media that are connected with these two previously disclosed and potentially state-backed operations on our service. We are making this ***data available*** *with the goal of encouraging open research and investigation of these behaviors from researchers and academics around the world. These large datasets comprise 3,841 accounts affiliated with the IRA, originating in Russia, and 770 other accounts, potentially originating in Iran. They include more than 10 million Tweets and more than 2 million images, GIFs, videos, and Periscope broadcasts."*

In June 2019, Twitter stated: "...we employ a range of open-source and proprietary signals and tools to identify when attempted coordinated manipulation may be taking place, as well as the actors responsible for it." Twitter blocking occurs *after* a significant period of unchecked online activity, which is what we analyze.

The Minority is making public an additional 1,103 accounts that were identified by Twitter subsequent to the November 1, 2017 hearing as connected to the IRA. Twitter has also informed that it removed 14 handles from the original list provided to Congress last fall, yielding an updated total of 3,841 Twitter accounts affiliated with the IRA. Twitter now believes those 14 accounts

[43] Marwick, A. E., & boyd, danah. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, *13*(1), 114–133. https://doi.org/10.1177/1461444810365313

should not be included based on improved methodology, improved understanding of IRA characteristics, and other new information – including the possibility that some are authentic user accounts that had become compromised.

Russian interference in the 2016 US presidential election led to multiple federal and industry investigations to identify malign actors and analyze their behavior As a part of these efforts, Twitter officially released a new dataset of 3,841 accounts believed to be connected to the Internet Research Agency (IRA). This dataset contains features such as profile description, account creation date, and poll choices. In our paper, we used the Russian troll accounts from this new dataset for our analysis, and model construction.

To conclude, Twitter released the screenames of almost 3.000 Twitter accounts that are believed to be connected to Russia's Internet Research Agency, after immediately deleting their Data from Twitter.com and Twitter API. A team at NBC News including *Ben Popken and EJ Fox* was able to reconstruct a dataset consisting of a subset of the deleted data for their investigation to figure how these troll accounts went on attack during key election moments. **This dataset is the body of this open-sourced reconstruction.**

Content

The dataset we used contains two CSV files, one called tweets.csv that includes details on individual tweets and one called users.csv, which includes details on individual accounts.

To recreate a link to a specific individual tweet found in the aforementioned dataset we replace user_key in https://twitter.com/user_key/status/tweet_id with the screen name from the user_key field and tweet_id with the number in the tweet_id field

If the links of this dataset are to be followed, they will lead to a suspended page on Twitter. However, there are some copies of the tweets as they originally appeared, including images, that can be found by entering the links on caches like archive.org and achive.is

Acknowledgements

Credits for the dataset used for this paper must be given to NBC News[44] that provided the public with it and to the US Senate's House Intelligence Committee

[44] https://www.nbcnews.com/tech/social-media/now-available-more-200-000-deleted-russian-troll-tweets-n844731

3.2 Feature Selection

As we mentioned above, the dataset we use in this project includes two files, one called *tweets.csv* and another one called *users.csv*.

Modeling of Troll Accounts

Features of tweets.csv

Regarding the features of the dataset file tweets.csv, we have a collection of 16 features that provide important utilizable information regarding the Russian trolls tweets. These features are:

- user_id
- user_key
- created_at
- created_str
- retweet_count
- retweeted
- favorite_count
- text
- tweet_id
- source, the tweet was posted as an HTML-formatted string. Tweets from the Twitter website have a source value of the web.
- hashtags
- expanded_urls
- posted
- mentioned
- in_reply_to_status_id

Features of users.csv

The dataset includes a sample of 14 features regarding the features of the users, that were marked as the Russian Troll Accounts. The features we used to visualize and extract information are:

- id
- location
- name
- followers_count
- statuses_count
- time_zone
- verified

- ➔ lang (language)
- ➔ screen_name
- ➔ description
- ➔ created_at
- ➔ favorites_count
- ➔ friends_count
- ➔ listed_count, which refers to the number of public lists that this user is a member of.

Timeline Analysis

Features of tweets.csv

Regarding the features of the dataset file tweets.csv, we have a collection of 16 features that provide important utilizable information regarding the Russian trolls tweets. These features are:

- ➔ user_id
- ➔ user_key
- ➔ created_at
- ➔ created_str
- ➔ retweet_count
- ➔ retweeted
- ➔ favorite_count
- ➔ text
- ➔ tweet_id
- ➔ source
- ➔ hashtags
- ➔ expanded_urls
- ➔ posted
- ➔ mentions
- ➔ retweeted_status_id
- ➔ in_reply_to_status_id

Features of users.csv

The dataset includes a sample of 14 features regarding the features of the users, that were marked as the Russian Troll Accounts. The features we used to visualize and extract information are:

- ➔ id
- ➔ location
- ➔ name
- ➔ followers_count
- ➔ statuses_count

- ➔ time_zone
- ➔ verified
- ➔ lang
- ➔ screen_name
- ➔ description
- ➔ created_at
- ➔ favourites_count
- ➔ friends_count
- ➔ listed_count

3.3 Programming Languages and Main Libraries

Python

Python is a popular general-purpose, interpreted, interactive, object-oriented, and high-level programming language, which is dynamically typed and garbage collected, meaning . Python was created by Guido van Rossum during the 1985-1990 period and released in 1991. Like pearl, python source code is also available under the GNU General Public License (GPL). Python supports a plethora of programming paradigms including Procedural , Object oriented and functional programming language. Its philosophy emphasized code readability with the use of major indentation.

Some of the main libraries we imported and used to implement our project are:

Pandas

Pandas is a Python package providing fast, flexible, and expressive data structures designed to make dealing with "relational" or "labeled" data both easy and intuitive. This package aims to be the fundamental high-level building block for doing practical, real world data analysis in Python. Furthermore, it has the bigger goal of becoming the most powerful and flexible open source data analysis manipulation tool available in any language.

Pandas_Profiling

Pandas profiling is an open source Python module which we use to conduct exploratory data analysis with a minimum number of code lines. Pandas profiling also generates interactive reports in web format that could be easily explainable to anyone regardless of coding knowledge. Pandas profiling generates a report with all the information ,retrieved from the dataset, that is easily available meaning it makes visualizing and understanding the distribution of each variable easier.

Pandas_profiling generates profile reports from a pandas DataFrame. Pandas profiling acts as an extension of pandas DataFrame by automatically generating a standardized univariate and multivariate report for data understanding.

According to Simon Brugman: "For each column, the following information[45] (whenever relevant for the column type) is presented in an interactive HTML report:

- Type inference: detect the types of columns in a DataFrame
- Essentials: type, unique values, indication of missing values
- Quantile statistics: minimum value, Q1, median, Q3, maximum, range, interquartile range
- Descriptive statistics: mean, mode, standard deviation, sum, median absolute deviation, coefficient of variation, kurtosis, skewness
- Most frequent and extreme values
- Histograms: categorical and numerical
- Correlations: high correlation warnings, based on different correlation metrics (Spearman, Pearson, Kendall, Cramér's V, Phik)
- Missing values: through counts, matrix, heatmap and dendrograms
- Duplicate rows: list of the most common duplicated rows
- Text analysis: most common categories (uppercase, lowercase, separator), scripts (Latin, Cyrillic) and blocks (ASCII, Cyrilic)
- File and Image analysis: file sizes, creation dates, dimensions, indication of truncated images and existance of EXIF metadata

The report contains three additional sections:

- Overview: mostly global details about the dataset (number of records, number of variables, overall missigness and duplicates, memory footprint)

[45] https://pandas-profiling.ydata.ai/docs/master/index.html

- Alerts: a comprehensive and automatic list of potential data quality issues (high correlation, skewness, uniformity, zeros, missing values, constant values, between others)
- Reproduction: technical details about the analysis (time, version and configuration)”

Seaborn

Seaborn is a library generically used for creating statistical graphics in Python. It is a python visualization library that is based on matplotlib and provides a high-level interface for forming inviting statistical graphics and integrates closely with *pandas* data structures.

Seaborn helps the researcher explore and understand more thoroughly the data that is being used. Its plotting functions operate on datagrames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce plots that provide useful information.

It has a dataset-orientated, declarative API that lets the user focus on the essence and understanding of the plots’ elements rather than the technical details regarding how to draw them.

Numpy

Numpy[46] which stands for Numerical Python, is a python library that is used for working with arrays and includes functions for working in domains of linear algebra, fourier transform and matrices. It is an open source project and is freely available to everyone.

In python, the purpose of arrays can be satisfied with the use of lists. However, they are slow to process, unlike Numpy that provides an array object up to fifty times faster than the traditional python lists. In NumPy the array object is called ndarray and it provides a plethora of supporting functions that make the utilization of ndarray very easy.

In data science, when speed and resources are of crucial importance, arrays are used. Numpy arrays are stored at one continuous place in memory unlike lists. As a result, processes can access and manipulate them very efficiently.

[46] https://www.w3schools.com/python/numpy/numpy_intro.asp

In computer science, the aforementioned trait is called locality, and is the main reason why NumPy is faster than lists. Numpy also is optimized to work with the latest CPU architectures.

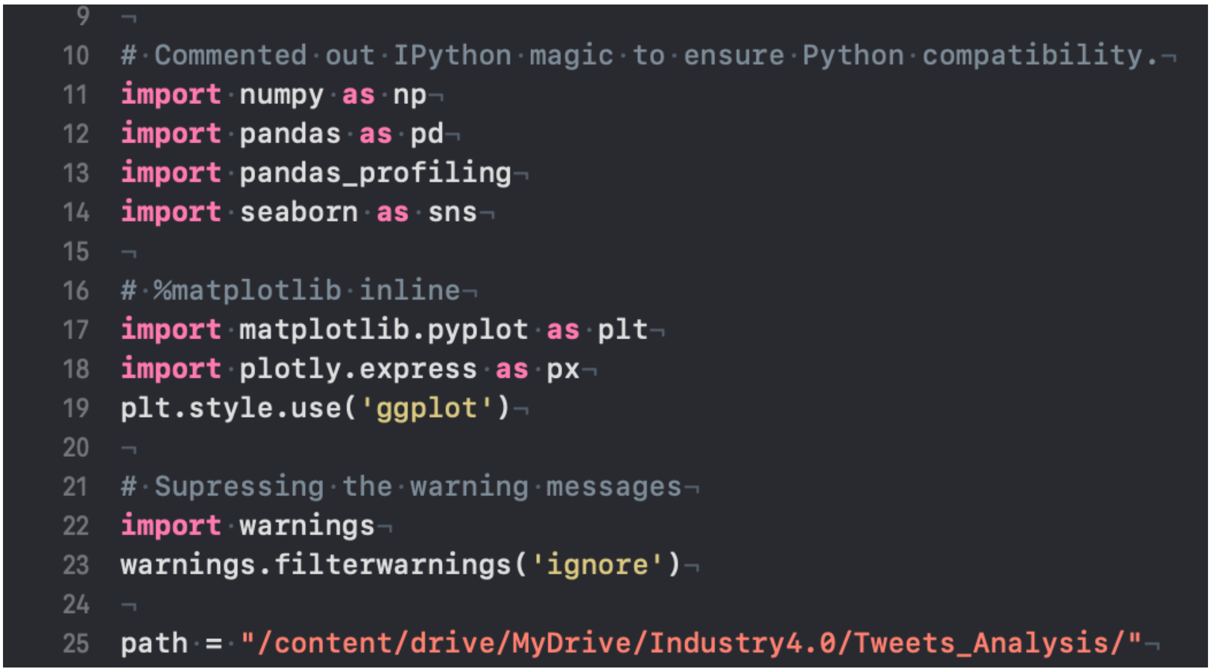
Chapter 4 Implementation & Results

4.1 Implementation

4.1.1 Modeling of Russian Troll Accounts

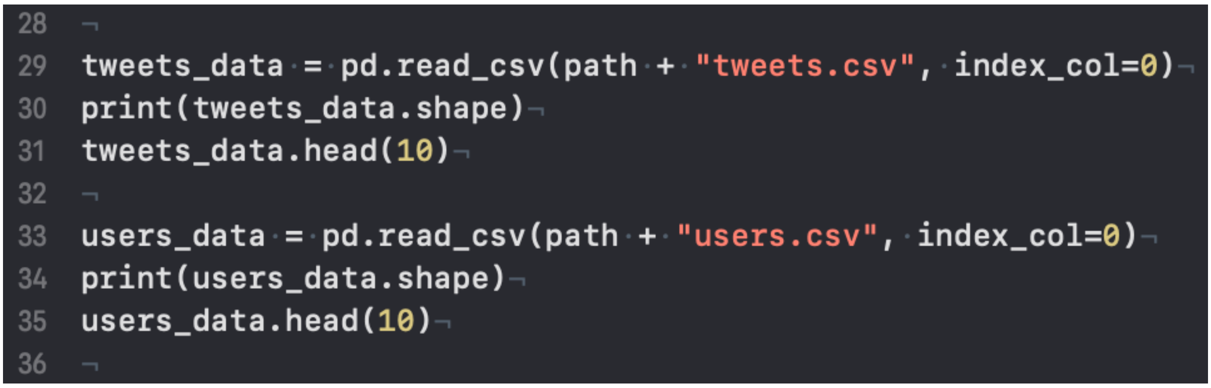
Importing Required Libraries

We start by Loading the Required Libraries, whose meanings we have explained thoroughly in the methodology chapter.



```

# Commented out IPython magic to ensure Python compatibility.
import numpy as np
import pandas as pd
import pandas_profiling
import seaborn as sns

# %matplotlib inline
import matplotlib.pyplot as plt
import plotly.express as px
plt.style.use('ggplot')

# Supressing the warning messages
import warnings
warnings.filterwarnings('ignore')

path = "/content/drive/MyDrive/Industry4.0/Tweets_Analysis/"
```

Importing Data



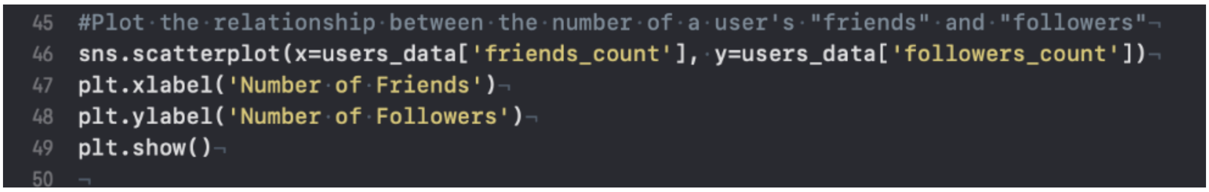
```

tweets_data = pd.read_csv(path + "tweets.csv", index_col=0)
print(tweets_data.shape)
tweets_data.head(10)

users_data = pd.read_csv(path + "users.csv", index_col=0)
print(users_data.shape)
users_data.head(10)

```

Following the preparative work of importing the required libraries and data, we proceed to the main purpose of the first part of this project which is to analyze and model characteristics of the Russian Troll accounts and tweets and conceptualize a first pattern of behavior and action.

Below, we attempt to answer crucial questions regarding the trolls' traits by interpreting output in a visualized form.

1. Are the users that the Russian trolls impersonate 'popular' ?

Input



```
#Plot the relationship between the number of a user's "friends" and "followers"
sns.scatterplot(x=users_data['friends_count'], y=users_data['followers_count'])
plt.xlabel('Number of Friends')
plt.ylabel('Number of Followers')
plt.show()
```

Output

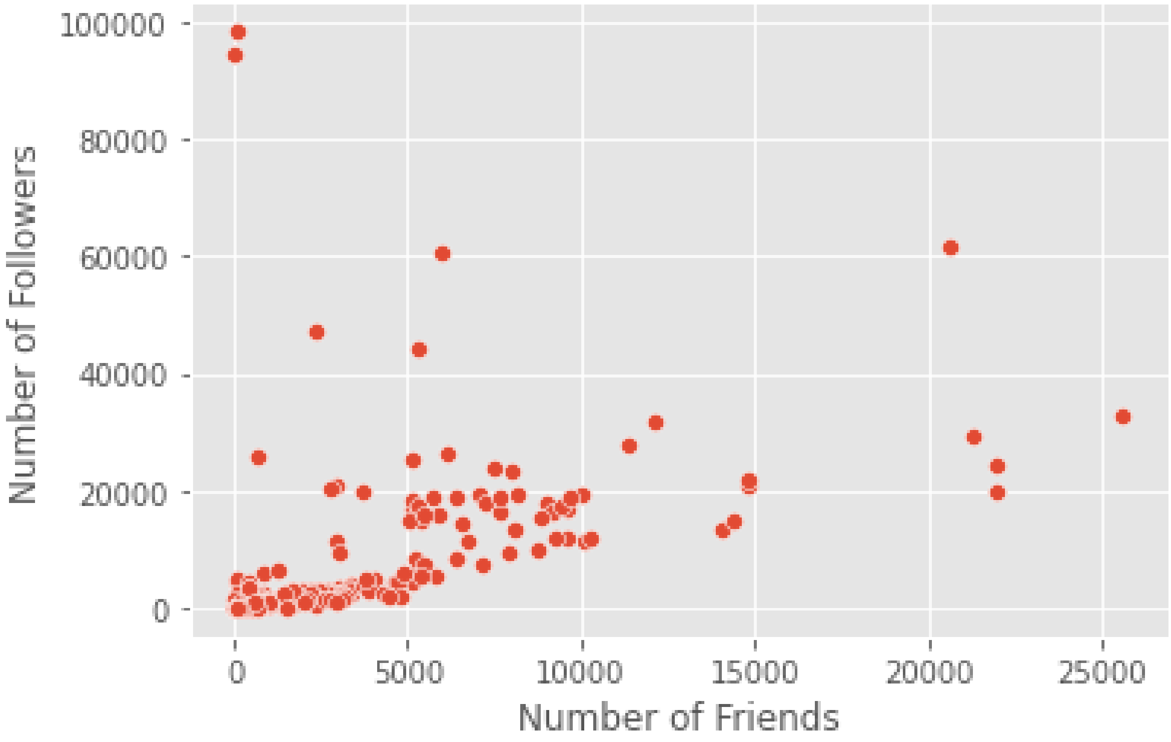


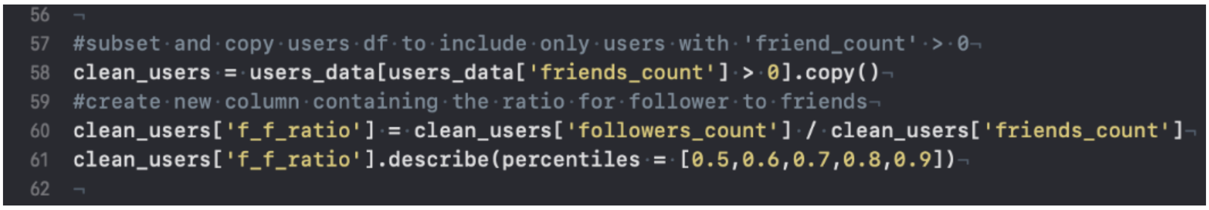
Figure 4 Scatter Plot showcasing the number of troll accounts' friends and follower counts

Remarks

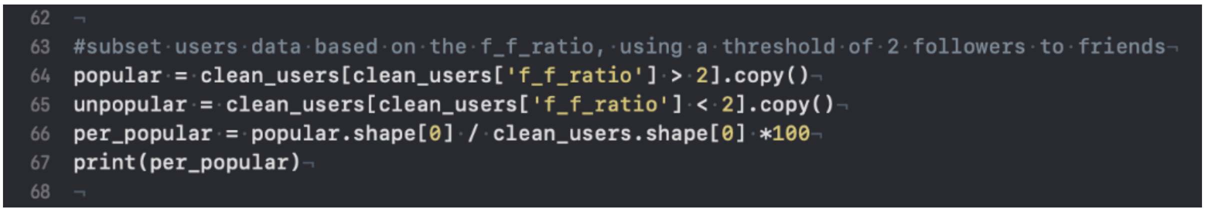
Firstly, it is of crucial importance to highlight the high number of followers & friends these troll accounts have, this is important because in unsuspected users' minds high numbers of followers are correlated with real accounts.

The above plot indicates a clear linear relationship for most users where the number of followers is ~ equal to the number of friends. That means that those accounts are followed back by the people they follow at an almost perfect rate. There are a few outliers in the top left of the scatter plot that do not comply with the pattern shared above. **Let's take a closer look at those users.**

Input:

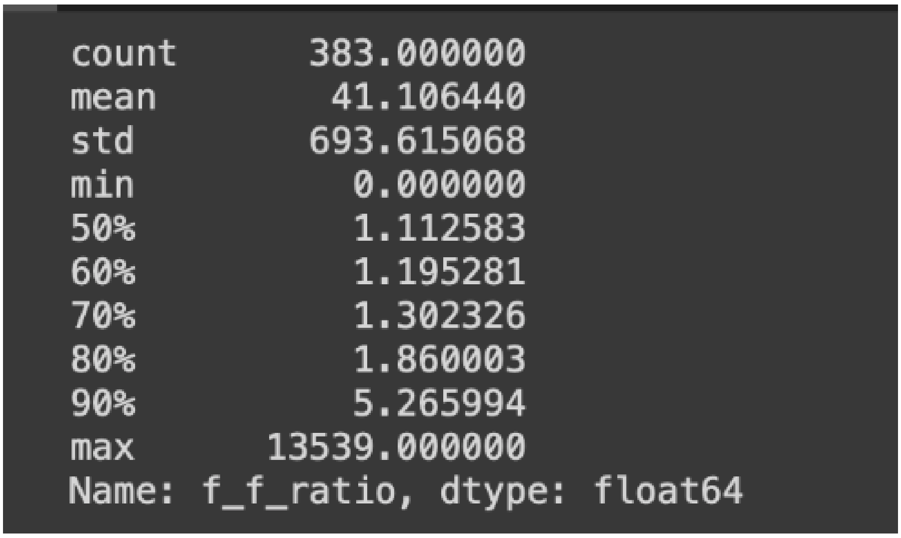


```

#subset and copy users df to include only users with 'friend_count' > 0
clean_users = users_data[users_data['friends_count'] > 0].copy()
#create new column containing the ratio for follower to friends
clean_users['f_f_ratio'] = clean_users['followers_count'] / clean_users['friends_count']
clean_users['f_f_ratio'].describe(percentiles = [0.5,0.6,0.7,0.8,0.9])

```



```

#subset users data based on the f_f_ratio, using a threshold of 2 followers to friends
popular = clean_users[clean_users['f_f_ratio'] > 2].copy()
unpopular = clean_users[clean_users['f_f_ratio'] < 2].copy()
per_popular = popular.shape[0] / clean_users.shape[0] *100
print(per_popular)

```

Output:

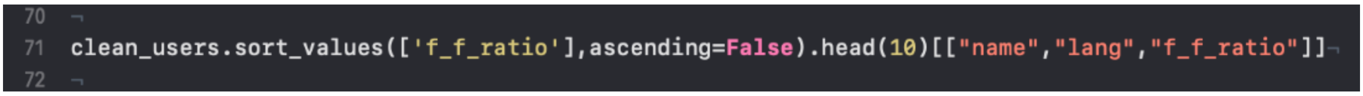


```
count       383.000000
mean         41.106440
std         693.615068
min           0.000000
50%           1.112583
60%           1.195281
70%           1.302326
80%           1.860003
90%           5.265994
max       13539.000000
Name: f_f_ratio, dtype: float64
```

Remarks:

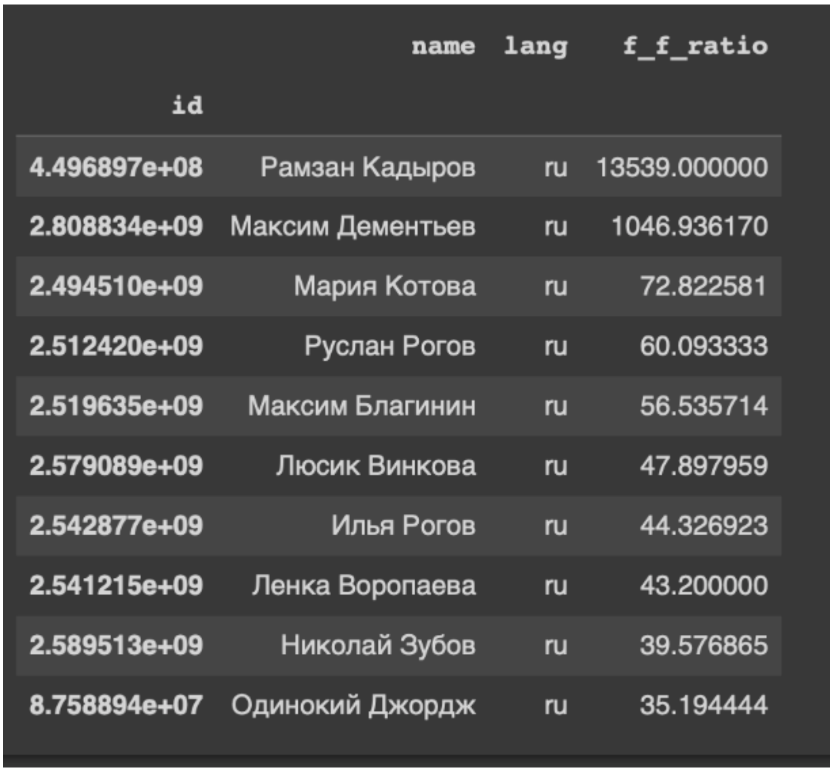
19% of these accounts have twice as many followers than friends - meaning at a twofold rate. **Who are these accounts that are so popular?**

Input:



```
clean_users.sort_values(['f_f_ratio'],ascending=False).head(10)[["name","lang","f_f_ratio"]]
```

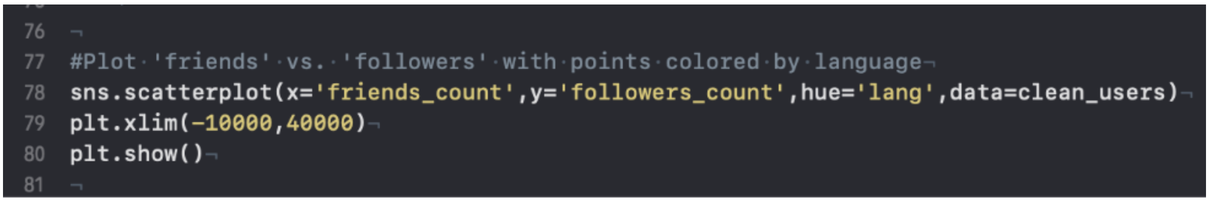
Output:



	name	lang	f_f_ratio
id			
4.496897e+08	Рамзан Кадыров	ru	13539.000000
2.808834e+09	Максим Дементьев	ru	1046.936170
2.494510e+09	Мария Котова	ru	72.822581
2.512420e+09	Руслан Рогов	ru	60.093333
2.519635e+09	Максим Благинин	ru	56.535714
2.579089e+09	Люсик Винкова	ru	47.897959
2.542877e+09	Илья Рогов	ru	44.326923
2.541215e+09	Ленка Воропаева	ru	43.200000
2.589513e+09	Николай Зубов	ru	39.576865
8.758894e+07	Одинокий Джордж	ru	35.194444

Remarks:
We can conclude from the above id diagram and data cleaning, that these popular users are all Russian, with a Russian name and writing in the Russian language, contrary to other trolls that mimic American citizens. **To validate this first indication, we move on to a more thorough data cleaning visualization.**

Input:



```

#Plot 'friends' vs. 'followers' with points colored by language
sns.scatterplot(x='friends_count',y='followers_count',hue='lang',data=clean_users)
plt.xlim(-10000,40000)
plt.show()

```

Output:

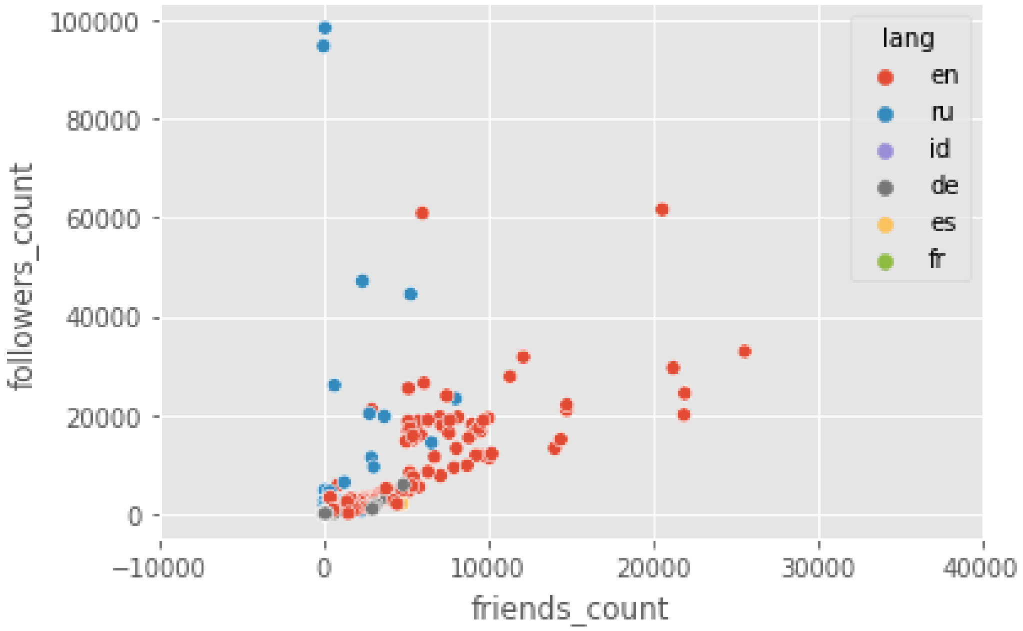


Figure 5 Scatter Plot showcasing the number of troll accounts' friends and follower counts and the language they are tweeting in

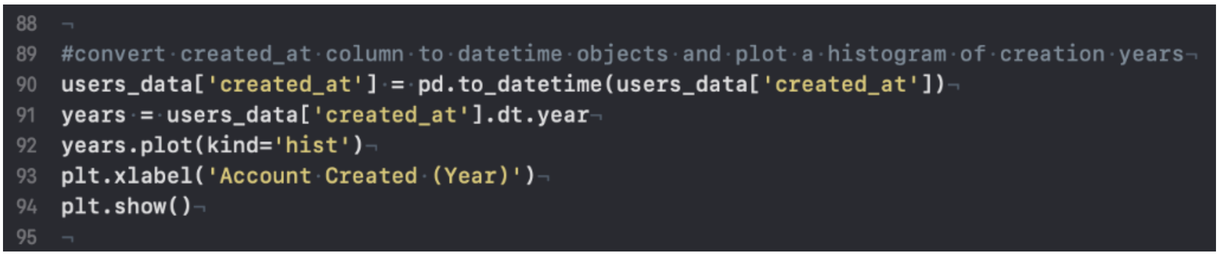
Remarks:

This plot confirms that the vast majority of the most popular accounts are troll accounts that use the Russian language, in comparison with accounts that use the English one.

2. When were the accounts created ?

That is a crucial indicator for the desired conclusion of this analysis, as the majority of these troll accounts were used to disturb and interfere in the US 2016 Presidential Elections. Consequently, if the time period they were created was prior or during the time the first campaign trails were formed and preliminary electoral procedures were taking place -- from early 2014 to late 2015 -- is clearly indicated that they were created for that sole purpose.

Input:



```

#convert created_at column to datetime objects and plot a histogram of creation years
users_data['created_at'] = pd.to_datetime(users_data['created_at'])
years = users_data['created_at'].dt.year
years.plot(kind='hist')
plt.xlabel('Account Created (Year)')
plt.show()

```

Output:

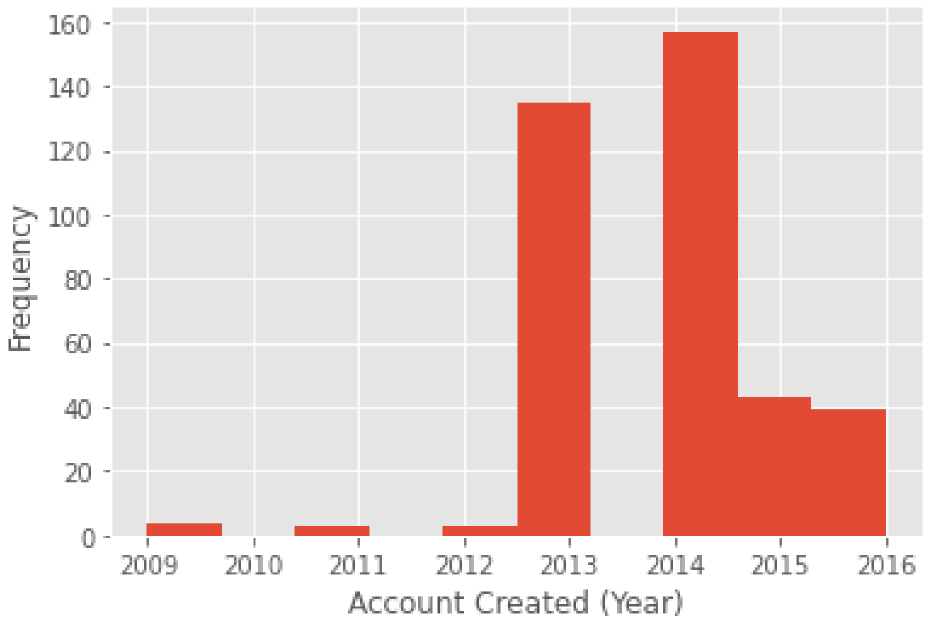
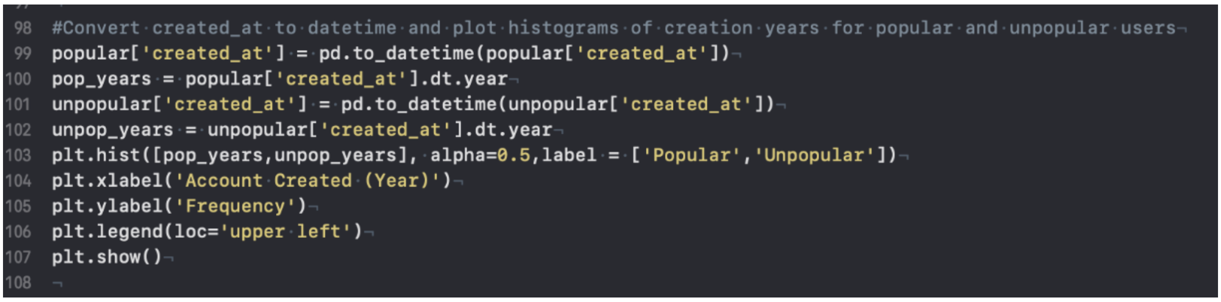


Figure 6 (Vertical) Bar Plot associating the year that troll accounts were created and at what frequency

Remarks:

Interestingly, as per the forenamed diagramma, the majority of the accounts were created in 2013, with the maximum creation reaching its peak in 2014, it's essential to note that accounts kept being created at a lower rate in 2015 and 2016. **When were the popular accounts created in comparison with the less popular accounts?**

Input:



```
#Convert created_at to datetime and plot histograms of creation years for popular and unpopular users
popular['created_at'] = pd.to_datetime(popular['created_at'])
pop_years = popular['created_at'].dt.year
unpopular['created_at'] = pd.to_datetime(unpopular['created_at'])
unpop_years = unpopular['created_at'].dt.year
plt.hist([pop_years,unpop_years], alpha=0.5,label = ['Popular','Unpopular'])
plt.xlabel('Account Created (Year)')
plt.ylabel('Frequency')
plt.legend(loc='upper left')
plt.show()

```

Output:

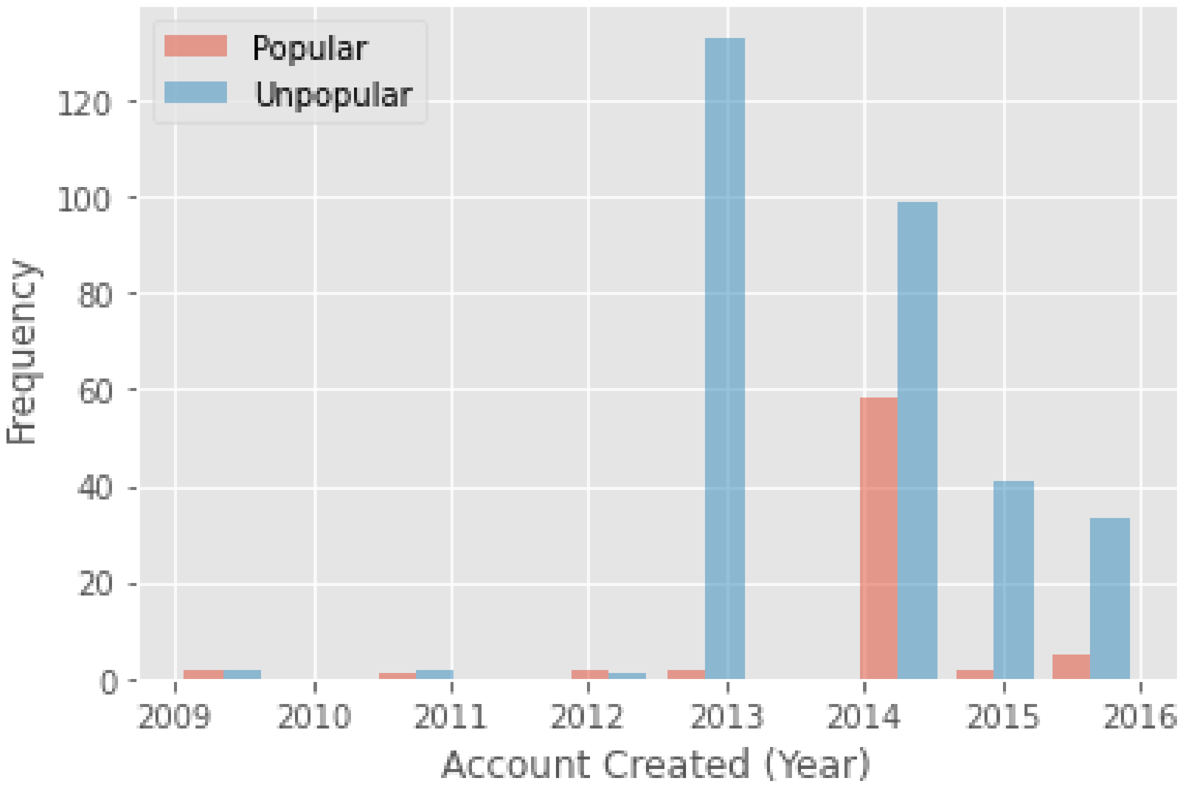
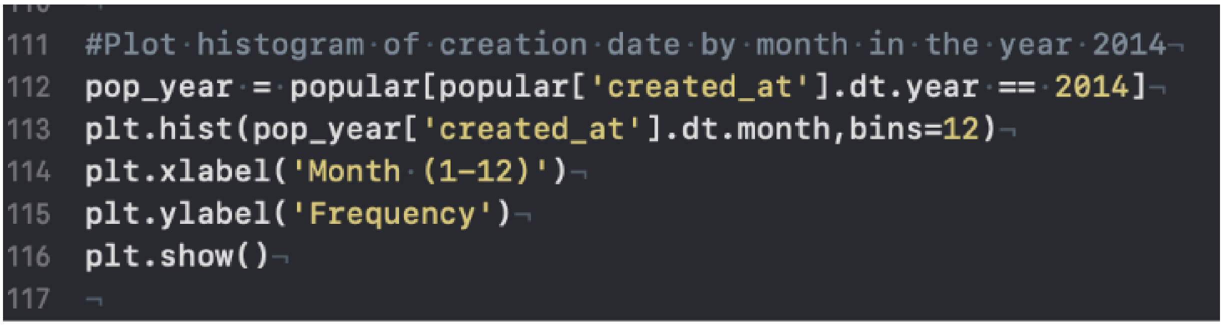


Figure 7 (Vertical) Grouped Bar Plot associating the year that the most popular troll accounts were created and at what frequency

Remarks:

Interestingly, the above graph indicates that the vast majority of the "popular" accounts were created in 2014, and the second most popular frequency can be seen in 2016. **Were all the popular accounts created at the same time in 2014?**

Input:



```
#Plot histogram of creation date by month in the year 2014
pop_year = popular[popular['created_at'].dt.year == 2014]
plt.hist(pop_year['created_at'].dt.month,bins=12)
plt.xlabel('Month (1-12)')
plt.ylabel('Frequency')
plt.show()
```

Output:

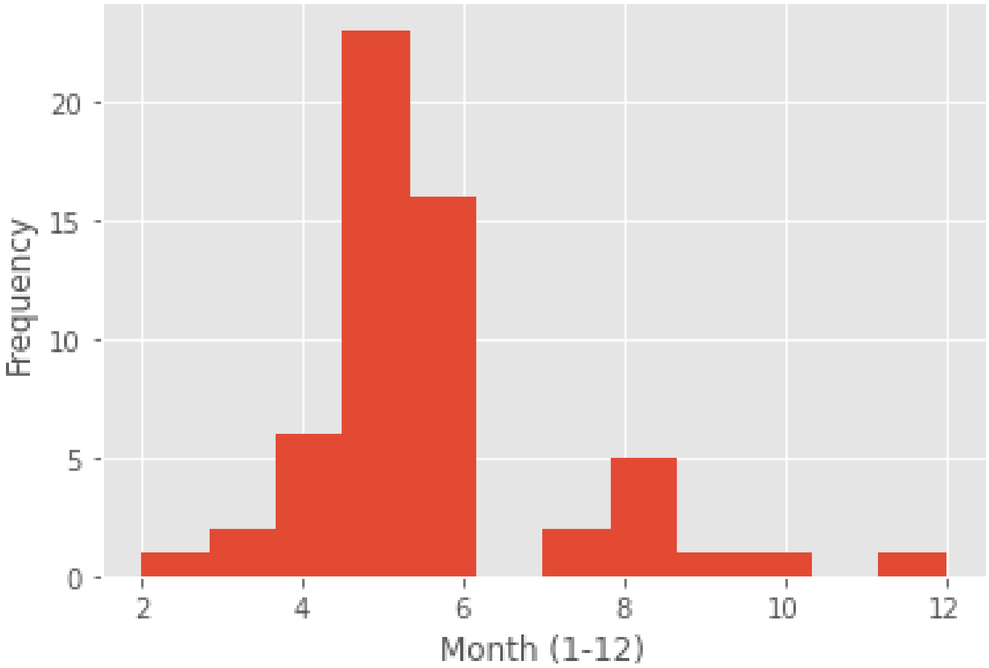


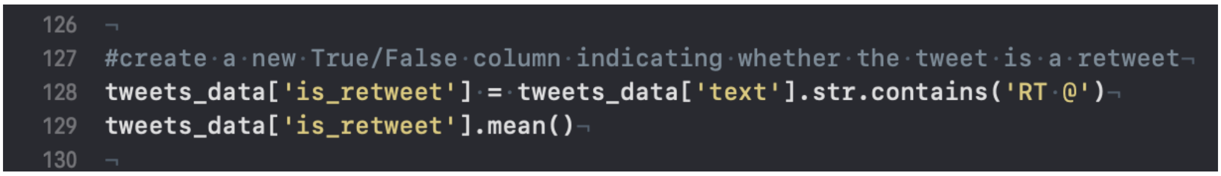
Figure 8 (Vertical) Bar Plot showcasing the month that the troll accounts were created in 2014 and at what frequency

Remarks:

Most of the popular accounts were created in May and June of 2014. That is really in accordance with the research made by the Senate's Intelligence Committee: "The field research to guide the attack appears to have begun **in earnest in June 2014.** Two Russian women, Aleksandra Y. Krylova and Anna V. Bogacheva, obtained visas for what turned out to be a three-week reconnaissance tour of the United States, including to key electoral states like Colorado, Michigan, Nevada and New Mexico...The two women bought cameras, SIM cards and disposable cell phones for the trip and devised "evacuation scenarios" in case their real purpose was detected. In all, they visited nine states — California, Illinois, Louisiana, New York and Texas, in addition to the others — "to gather intelligence" on American politics, the indictment says. Ms. Krylova sent a report about their findings to one of her bosses in St. Petersburg."

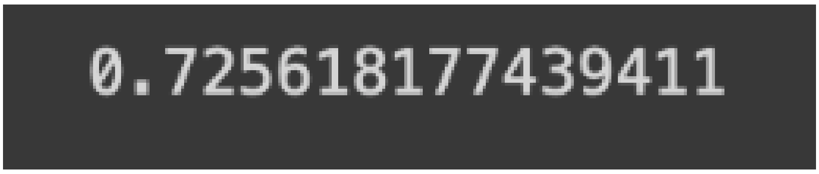
3. What fraction of the collected Russian Troll Tweets were actually retweets ?

Input:



```

#create a new True/False column indicating whether the tweet is a retweet
tweets_data['is_retweet'] = tweets_data['text'].str.contains('RT @')
tweets_data['is_retweet'].mean()

```

Output:

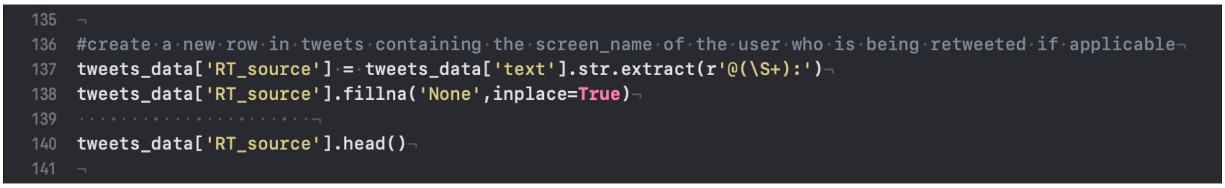


```
0.725618177439411
```

Remarks:

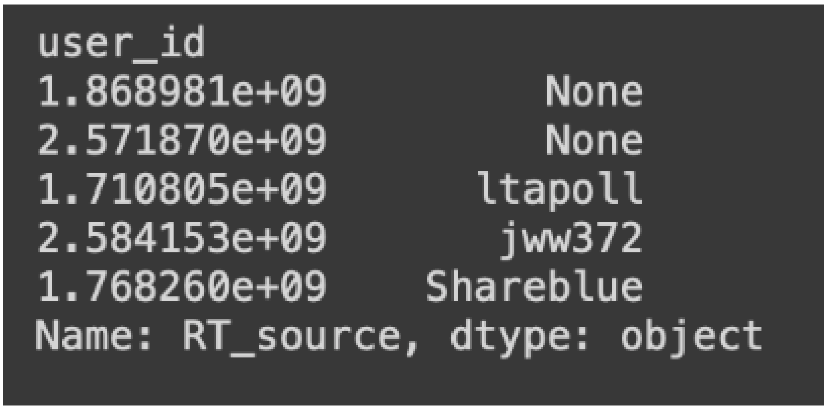
Almost 73% of the tweets, in the dataset, are retweets. **Who are the accounts that are getting retweeted? Are they Russian users or 'Real' users?**

Input:



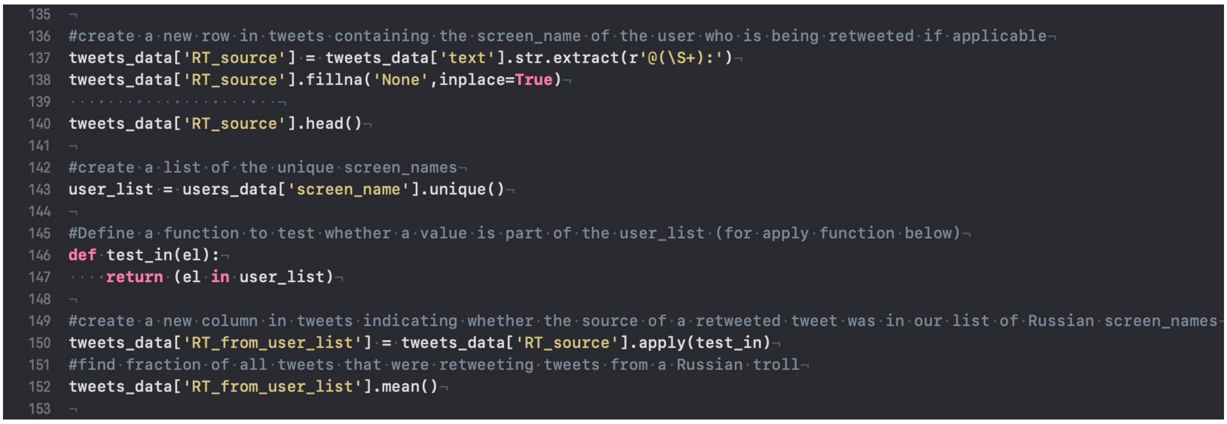
```

#create a new row in tweets containing the screen_name of the user who is being retweeted if applicable
tweets_data['RT_source'] = tweets_data['text'].str.extract(r'@(\S+):')
tweets_data['RT_source'].fillna('None',inplace=True)

tweets_data['RT_source'].head()

```

Output:



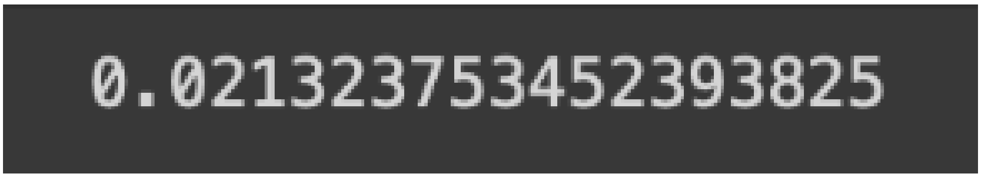
```
user_id
1.868981e+09          None
2.571870e+09          None
1.710805e+09       ltapoll
2.584153e+09        jww372
1.768260e+09     Shareblue
Name: RT_source, dtype: object
```

Input:



```
#create a new row in tweets containing the screen_name of the user who is being retweeted if applicable
tweets_data['RT_source'] = tweets_data['text'].str.extract(r'@(\S+):')
tweets_data['RT_source'].fillna('None',inplace=True)

tweets_data['RT_source'].head()

#create a list of the unique screen_names
user_list = users_data['screen_name'].unique()

#Define a function to test whether a value is part of the user_list (for apply function below)
def test_in(el):
    return (el in user_list)

#create a new column in tweets indicating whether the source of a retweeted tweet was in our list of Russian screen_names
tweets_data['RT_from_user_list'] = tweets_data['RT_source'].apply(test_in)
#find fraction of all tweets that were retweeting tweets from a Russian troll
tweets_data['RT_from_user_list'].mean()

```

Output:



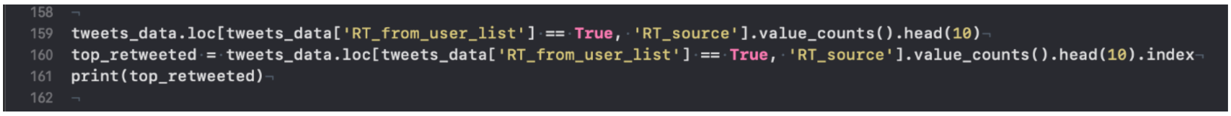
```
0.021323753452393825
```

Remarks:

Only about 2% of the retweeted tweets were originally composed by a Russian troll. In other words, a very small fraction of the retweets were amplifying an original idea composed by a Russian troll, they were rather prefering to retweet a tweet composed by an actual account, capitalizing on their legitimacy.

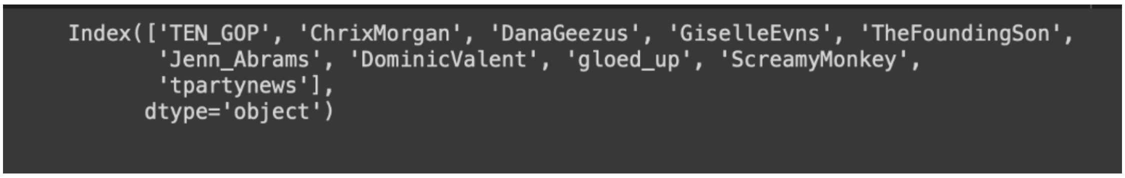
4. Which users from within the list of Russian trolls were most commonly retweeted ?

Input:



```

tweets_data.loc[tweets_data['RT_from_user_list'] == True, 'RT_source'].value_counts().head(10)
top_retweeted = tweets_data.loc[tweets_data['RT_from_user_list'] == True, 'RT_source'].value_counts().head(10).index
print(top_retweeted)

```

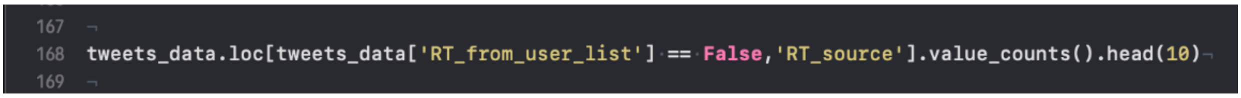
Output:



```
Index(['TEN_GOP', 'ChrixMorgan', 'DanaGeezus', 'GiselleEvns', 'TheFoundingSon',
       'Jenn_Abrams', 'DominicValent', 'gloed_up', 'ScreamyMonkey',
       'tpartynews'],
      dtype='object')
```

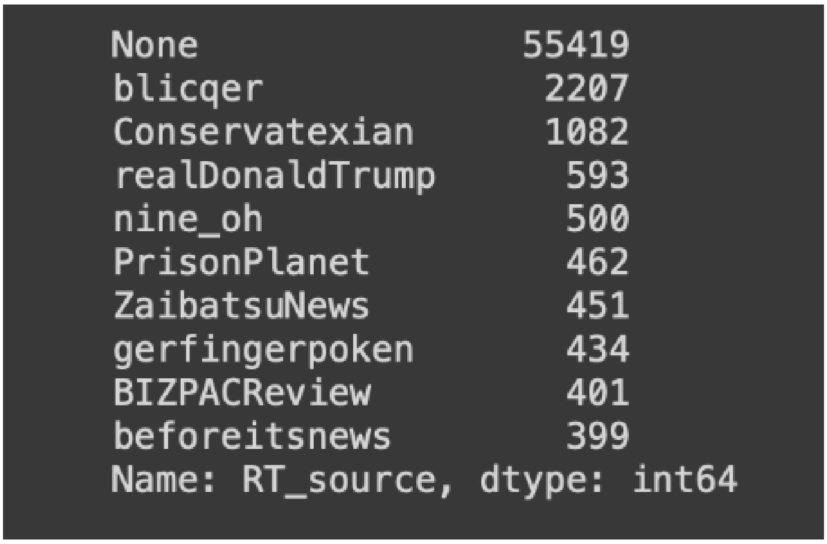
5. Who are the users who were most frequently retweeted outside of the Russian troll list?

Input:



```
tweets_data.loc[tweets_data['RT_from_user_list'] == False,'RT_source'].value_counts().head(10)
```

Output:



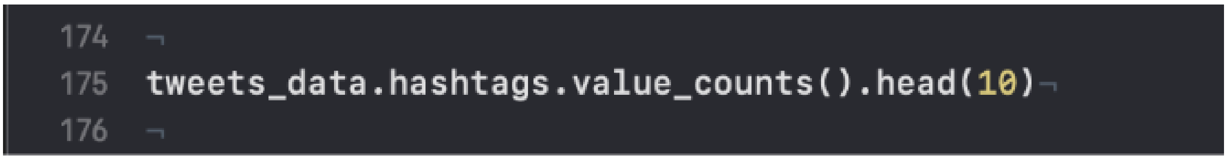
```
None                55419
blicqer              2207
Conservatexian       1082
realDonaldTrump       593
nine_oh               500
PrisonPlanet          462
ZaibatsuNews          451
gerfingerpoken        434
BIZPACReview          401
beforeitsnews         399
Name: RT_source, dtype: int64
```

Remarks:

This list, which is mainly Convservative-Republican accounts, and actual Donald Trump's tweets, supports the claim made by Senate's Intelligence committee that found that "the IRA sought to influence the 2016 U.S. presidential election by harming Hillary Clinton’s chances of success and supporting Donald Trump at the direction of the Kremlin. The Committee found that IRA social media activity was overtly and almost invariably supportive of then-candidate Trump to the detriment of Secretary Clinton’s campaign.”

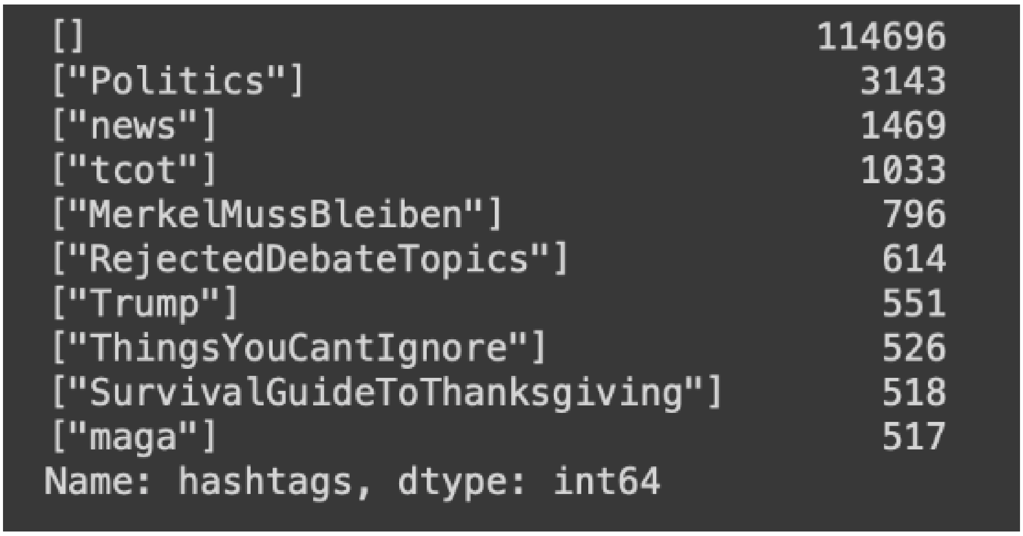
6. What are the most frequent hashtags used ?

Input:



```
tweets_data.hashtags.value_counts().head(10)
```

Output:



```
[]                                         114696
["Politics"]                                 3143
["news"]                                     1469
["tcot"]                                     1033
["MerkelMussBleiben"]                         796
["RejectedDebateTopics"]                      614
["Trump"]                                     551
["ThingsYouCantIgnore"]                       526
["SurvivalGuideToThanksgiving"]               518
["maga"]                                      517
Name: hashtags, dtype: int64
```

Remarks:

The list of most used hashtags, clearly indicates that the russian troll accounts were tweeting about political issues, the news and were supporting the Donald Trump & "Make America Great Again" Movement, while also perpetuating the fake news that circulated the internet at the time, regarding a Rejected Debate Topics Narrative, that implied that Hillary Rodham Clinton rejected Debate topics, due to incapability of answering questions about them.

Final Remarks of 4.1.1

Ultimately, in order to influence the mind's of Americans, Russian troll's must achieve reach and credibility. The first, reach, they are able to achieve through a large amount of followers. Through our modeling and analysis we have found that there is significant evidence pointing towards the idea that Russian bot's are following each other in order to gain followers, and thus gain reach. We have also found through content analysis of their tweets, that they are potentially using buzzwords to raise the promotion and thus reach of their tweets.

In addition, we learned through analysis of the publish date of these tweets, that the trolls were not only active before the 2016 election, but they continued afterwards--signifying that their goal was not only to affect the 2016 election but to create broader political discourse in the United States.

One of the most surprising things we concluded was just how much reach these russian trolls have. Many of these trolls had tens of thousands of followers on twitter which in addition to giving them greater reach, give them greater validity, because people think that higher numbers of followers indicate that the account belongs to an actual person. Probably the most significant thing we learned, however, is just how well disguised these tweets are. Not all of them include "#LockHerUp" or obvious anti-democrat messages, instead they are more hidden. This can best be shown by the list with most used hashtags.

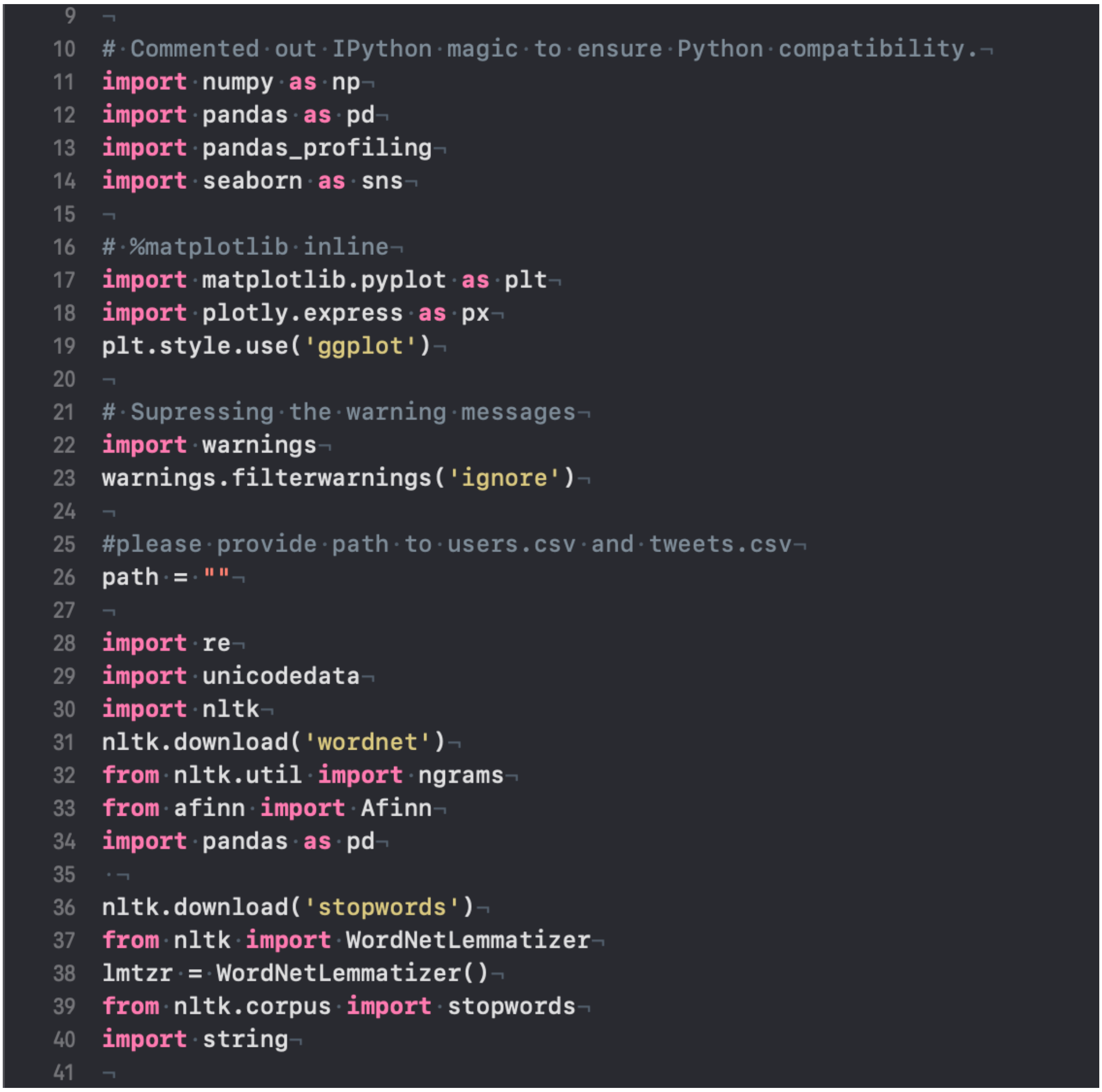
4.1.2 Timeline Analysis

Background

As part of the House Intelligence Committee investigation into how Russia may have influenced the 2016 US Election, Twitter released the screen names of almost 3000 Twitter accounts believed to be connected to Russia's Internet Research Agency, a company known for operating social media troll accounts. The Mueller's investigation released on Friday, 16th February 2018 an indictment of Russian Operatives Details Playbook Of Information Warfare. (NPR). In the following attempt to get a better understanding and deep dive into how russian trolls have influenced the US 2016 Presidential elections we focus among others in producing and visualizing a timeline correlated with the 2016 election, correlation between accounts created, tweets text analysis, sentiment analysis and public opinion twists in accordance with specific events and tweets.

Importing Required Libraries

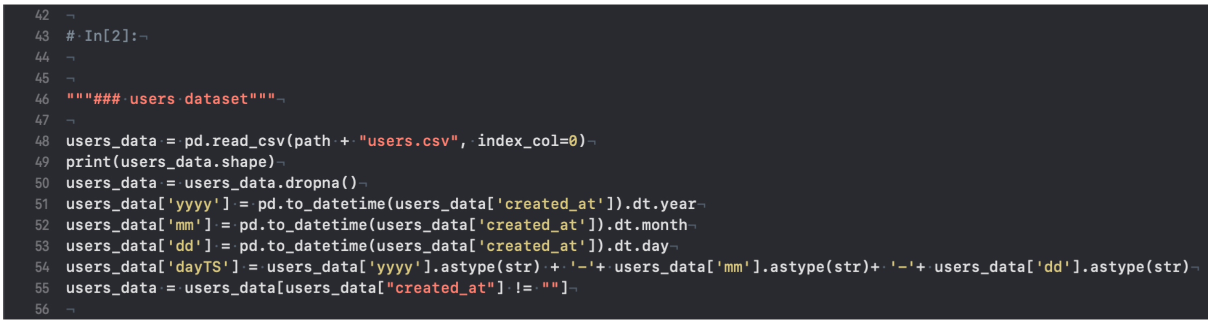
We start by Loading the Required Libraries



```python

# Commented out IPython magic to ensure Python compatibility.
import numpy as np
import pandas as pd
import pandas_profiling
import seaborn as sns

# %matplotlib inline
import matplotlib.pyplot as plt
import plotly.express as px
plt.style.use('ggplot')

# Supressing the warning messages
import warnings
warnings.filterwarnings('ignore')

#please provide path to users.csv and tweets.csv
path = ""

import re
import unicodedata
import nltk
nltk.download('wordnet')
from nltk.util import ngrams
from afinn import Afinn
import pandas as pd
 
nltk.download('stopwords')
from nltk import WordNetLemmatizer
lmtzr = WordNetLemmatizer()
from nltk.corpus import stopwords
import string

```

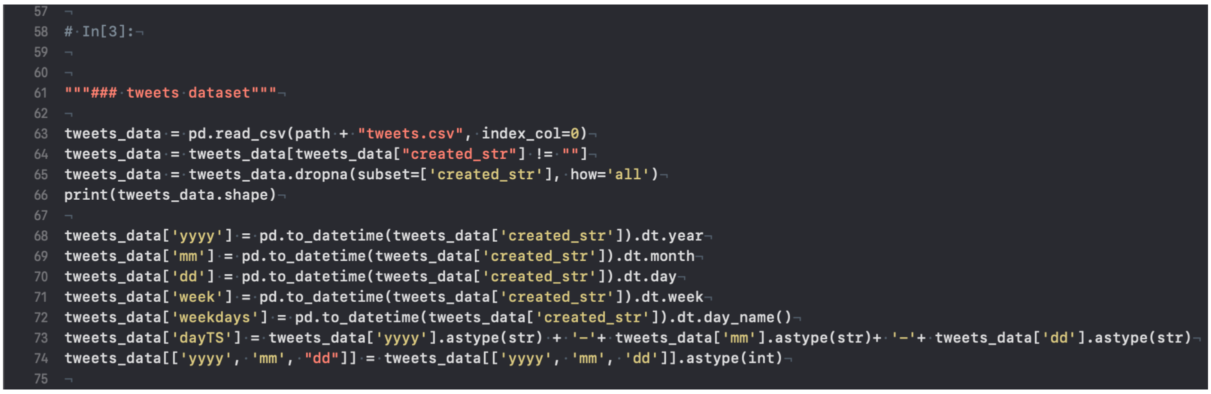
Data Preparation

In this section we load the Users Dataset and **Extract Year, Month, Day and Date** out of it



```python

# In[2]:


"""### users dataset"""

users_data = pd.read_csv(path + "users.csv", index_col=0)
print(users_data.shape)
users_data = users_data.dropna()
users_data['yyyy'] = pd.to_datetime(users_data['created_at']).dt.year
users_data['mm'] = pd.to_datetime(users_data['created_at']).dt.month
users_data['dd'] = pd.to_datetime(users_data['created_at']).dt.day
users_data['dayTS'] = users_data['yyyy'].astype(str) + '-'+ users_data['mm'].astype(str)+ '-'+ users_data['dd'].astype(str)
users_data = users_data[users_data["created_at"] != ""]

```

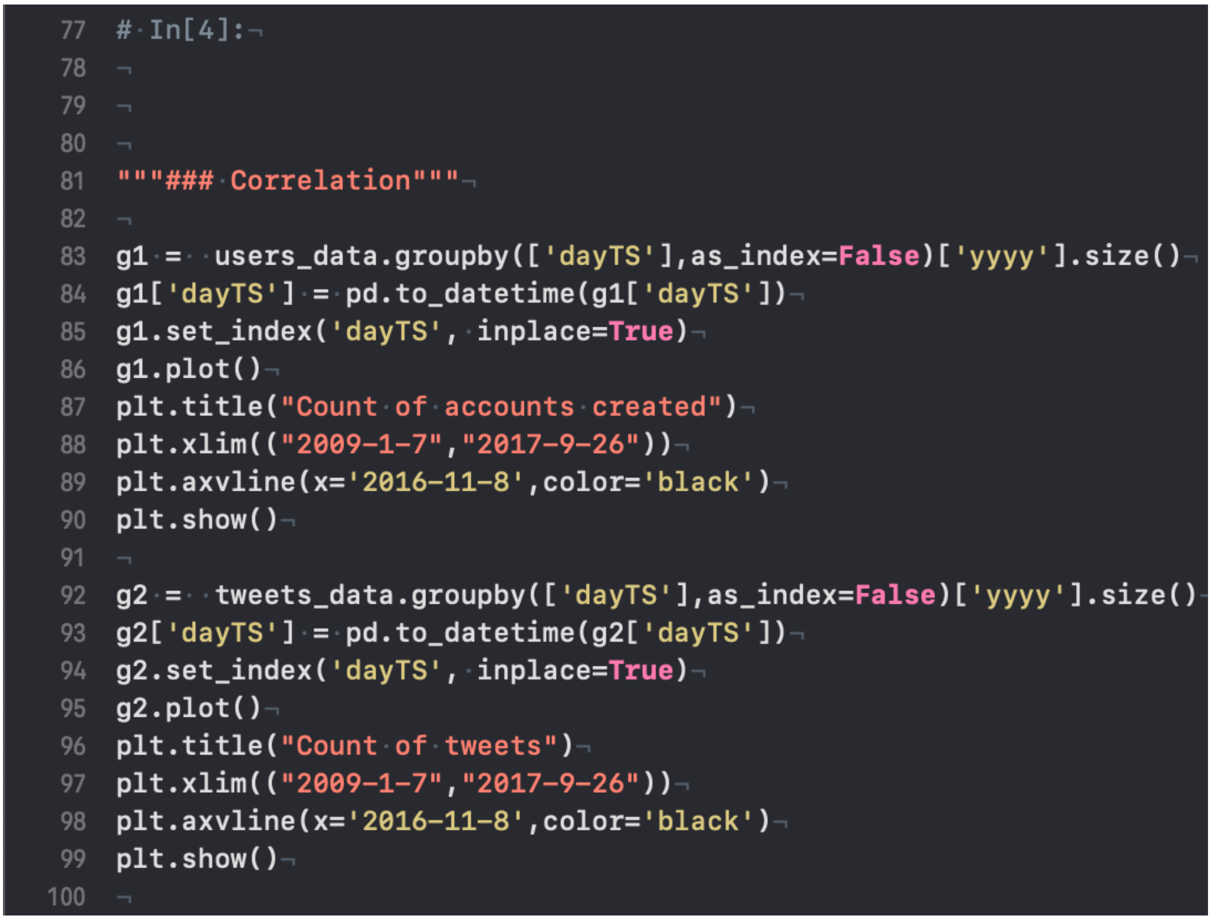
Below, we load the Tweets Dataset and Extract Year, Month, Day and Date out of it



```

# In[3]:


"""### tweets dataset"""

tweets_data = pd.read_csv(path + "tweets.csv", index_col=0)
tweets_data = tweets_data[tweets_data["created_str"] != ""]
tweets_data = tweets_data.dropna(subset=['created_str'], how='all')
print(tweets_data.shape)

tweets_data['yyyy'] = pd.to_datetime(tweets_data['created_str']).dt.year
tweets_data['mm'] = pd.to_datetime(tweets_data['created_str']).dt.month
tweets_data['dd'] = pd.to_datetime(tweets_data['created_str']).dt.day
tweets_data['week'] = pd.to_datetime(tweets_data['created_str']).dt.week
tweets_data['weekdays'] = pd.to_datetime(tweets_data['created_str']).dt.day_name()
tweets_data['dayTS'] = tweets_data['yyyy'].astype(str) + '-'+ tweets_data['mm'].astype(str)+ '-'+ tweets_data['dd'].astype(str)
tweets_data[['yyyy', 'mm', "dd"]] = tweets_data[['yyyy', 'mm', 'dd']].astype(int)

```

Correlation

Next step involves preparing the **Correlation plot** and showcasing correlation between the created account and tweets



```
# In[4]:



"""### Correlation"""

g1 =  users_data.groupby(['dayTS'],as_index=False)['yyyy'].size()
g1['dayTS'] = pd.to_datetime(g1['dayTS'])
g1.set_index('dayTS', inplace=True)
g1.plot()
plt.title("Count of accounts created")
plt.xlim(("2009-1-7","2017-9-26"))
plt.axvline(x='2016-11-8',color='black')
plt.show()

g2 =  tweets_data.groupby(['dayTS'],as_index=False)['yyyy'].size()
g2['dayTS'] = pd.to_datetime(g2['dayTS'])
g2.set_index('dayTS', inplace=True)
g2.plot()
plt.title("Count of tweets")
plt.xlim(("2009-1-7","2017-9-26"))
plt.axvline(x='2016-11-8',color='black')
plt.show()

```

Output:

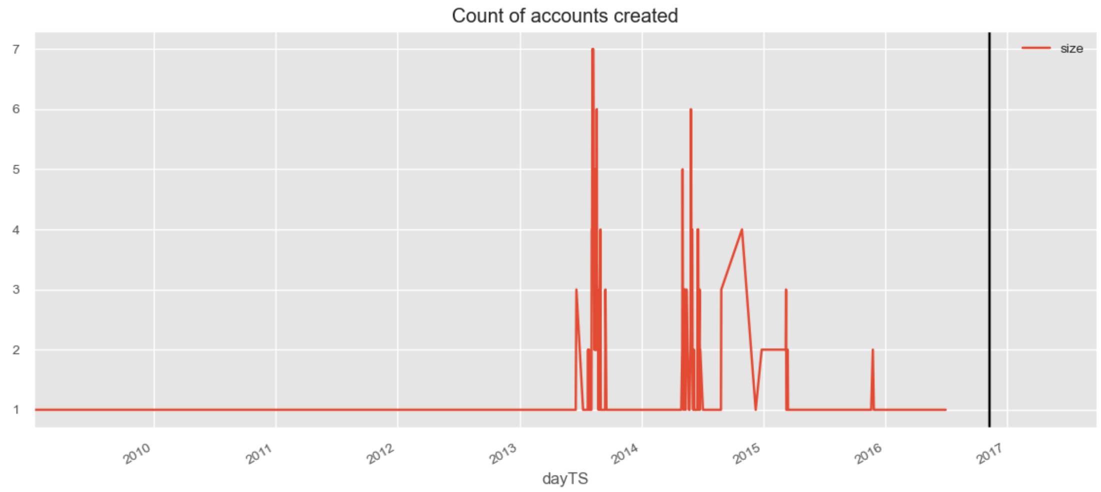


Figure 9 Hits Graph connecting the count of troll accounts created with the day,month and year they did.

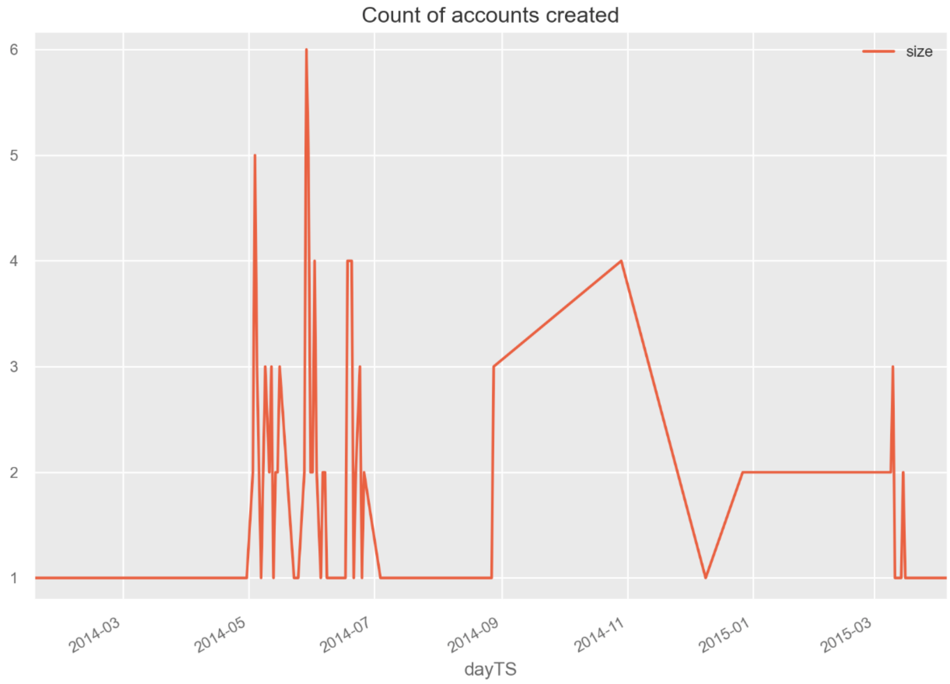


Figure 10 Zoomed in part of Figure 9 Hits Graph for the time period from March 2014 until March 2015

Remarks:

According to the data provided by the dataset we used, we can witness that despite the fact that the US presidential elections were scheduled to take place in 2016, the russian disinformation campaign started 3 years earlier, as the first russian troll account was created mid 2013. Furthermore, we can witness a rise in the count of troll accounts created during the period between **May 2014 and July 2014.**

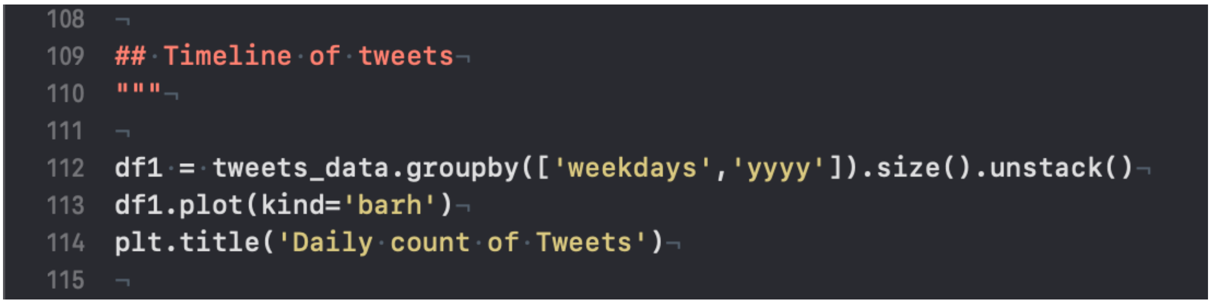
That is justified by the FBI analysis, posted in the New York Times, which states that "the field research to guide the attack appears to have begun in earnest in **June 2014.** Two Russian women, Aleksandra Y. Krylova and Anna V. Bogacheva, obtained visas for what turned out to be a three-week reconnaissance tour of the United States, including to key electoral states like Colorado, Michigan, Nevada and New Mexico. The visa application of a third Russian, Robert S. Bovda, was rejected."

The two women bought cameras, **SIM cards and disposable cell phones** for the trip and devised "evacuation scenarios" in case their real purpose was detected. In all, they visited nine states — California, Illinois, Louisiana, New York and Texas, in addition to the others — "to gather intelligence" on American politics, the indictment says. Ms. Krylova sent a report about their findings to one of her bosses in St. Petersburg.

Another Russian[47] operative visited Atlanta in **November 2014** on a similar mission, the indictment says. It does not name that operative, a possible indication that he or she is cooperating with the investigation, legal experts said.

The aforementioned incident provides a valuable link between the data we mined and visualized and actual real-life events that assisted the three year Russian disinformation campaign.

Plotting daily count of Tweets



```

## Timeline of tweets
"""

df1 = tweets_data.groupby(['weekdays','yyyy']).size().unstack()
df1.plot(kind='barh')
plt.title('Daily count of Tweets')

```

Output:

[47] https://www.nytimes.com/2018/02/16/us/politics/russia-mueller-election.html

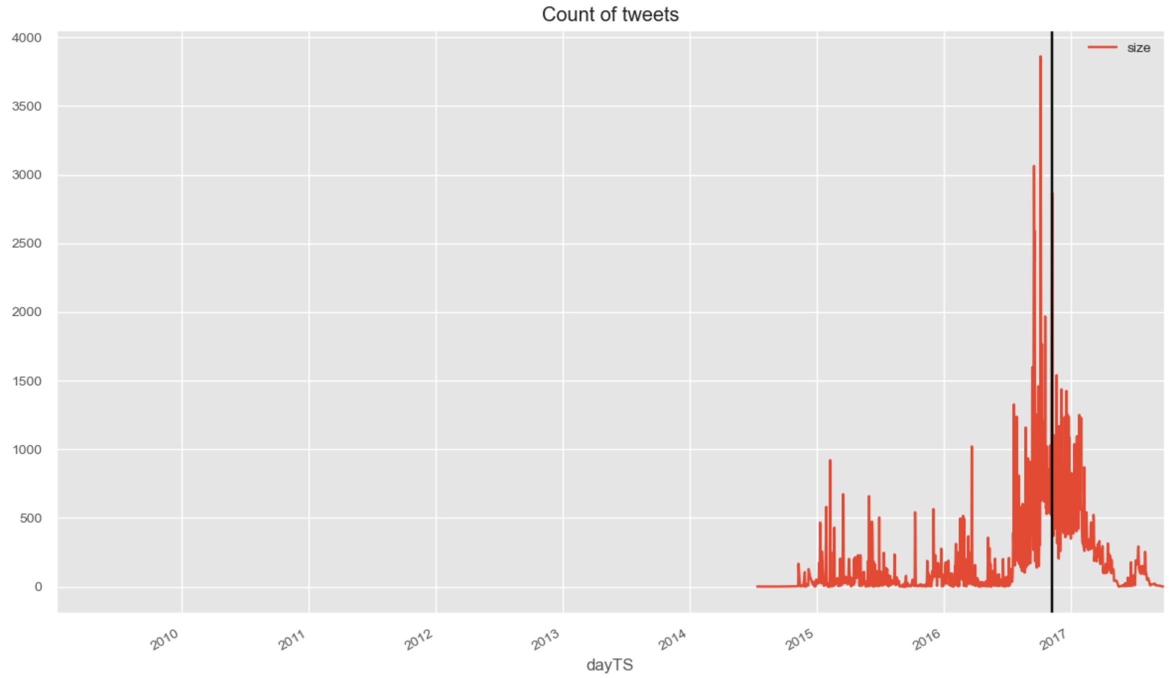
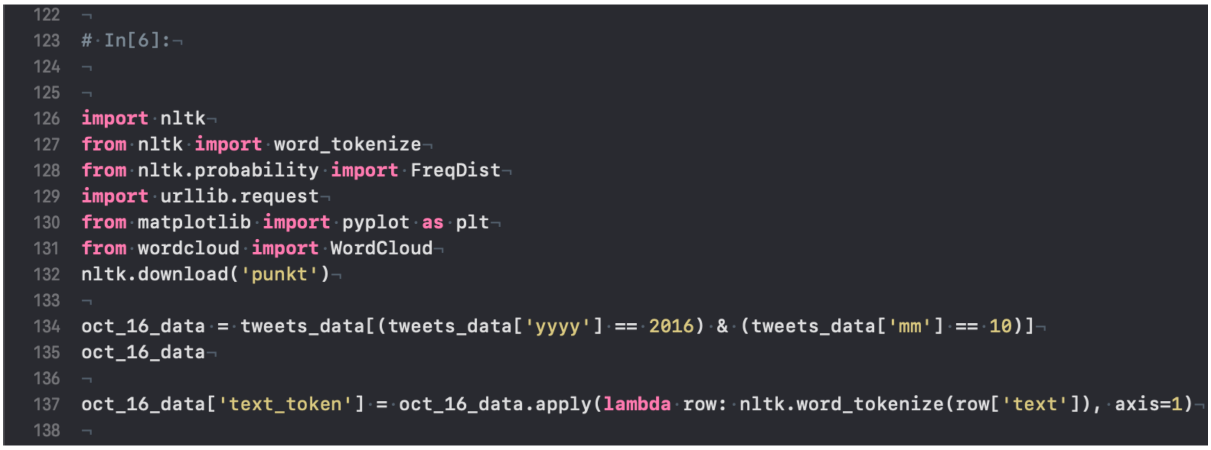


Figure 10 Hits Graph connecting the count of troll tweets with the day,month and year they were posted.

Closer Look at 2016

Filtering data to get October 2016 Data



```

# In[6]:


import nltk
from nltk import word_tokenize
from nltk.probability import FreqDist
import urllib.request
from matplotlib import pyplot as plt
from wordcloud import WordCloud
nltk.download('punkt')

oct_16_data = tweets_data[(tweets_data['yyyy'] == 2016) & (tweets_data['mm'] == 10)]
oct_16_data

oct_16_data['text_token'] = oct_16_data.apply(lambda row: nltk.word_tokenize(row['text']), axis=1)

```

Output:

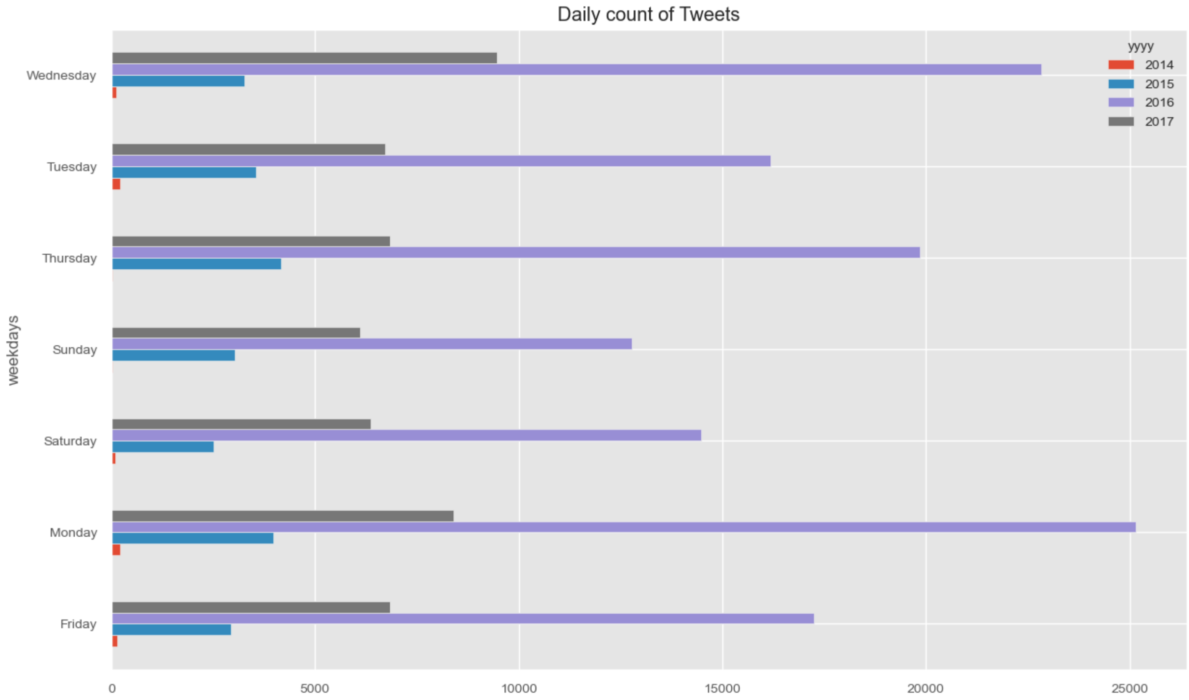
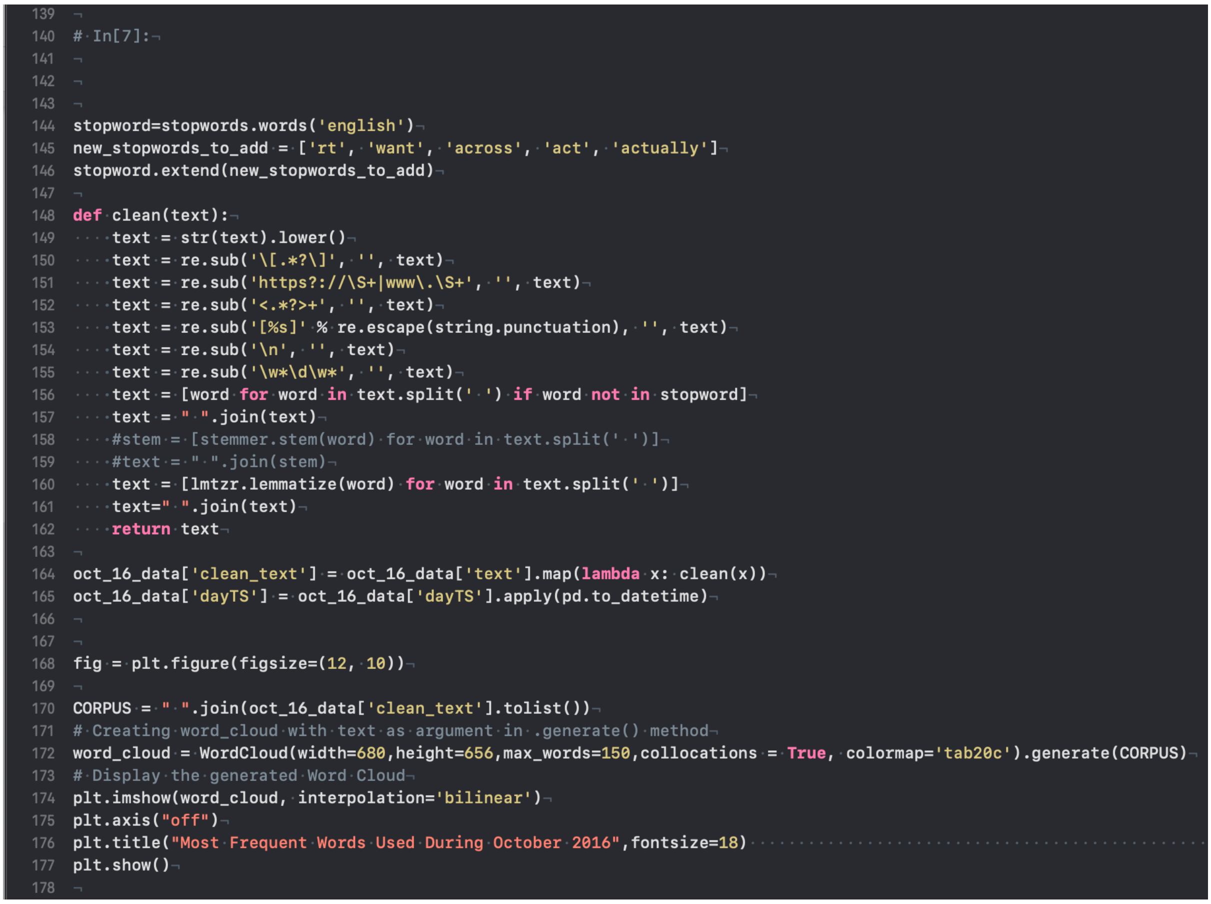


Figure 11 Horizontal Grouped Bar Plot correlating the daily count of tweets, weekday and year

Tweets Content Analysis

Creating a Function to clean the Text Column, the function takes in the text column converts the text to lowercase, removes unwanted and non english characters, stopwords and uses the concept of **Lemmatization** to get the base form of word, finally this function is applied to the text column and a new column called **‘clean_text’** is created which is then used to create a word cloud of most frequent Words used during October 2016



```

# In[7]:



stopword=stopwords.words('english')
new_stopwords_to_add = ['rt', 'want', 'across', 'act', 'actually']
stopword.extend(new_stopwords_to_add)

def clean(text):
    text = str(text).lower()
    text = re.sub('\[.*?\]', '', text)
    text = re.sub('https?://\S+|www\.\S+', '', text)
    text = re.sub('<.*?>+', '', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
    text = re.sub('\n', '', text)
    text = re.sub('\w*\d\w*', '', text)
    text = [word for word in text.split(' ') if word not in stopword]
    text = " ".join(text)
    #stem = [stemmer.stem(word) for word in text.split(' ')]
    #text = " ".join(stem)
    text = [lmtzr.lemmatize(word) for word in text.split(' ')]
    text=" ".join(text)
    return text

oct_16_data['clean_text'] = oct_16_data['text'].map(lambda x: clean(x))
oct_16_data['dayTS'] = oct_16_data['dayTS'].apply(pd.to_datetime)


fig = plt.figure(figsize=(12, 10))

CORPUS = " ".join(oct_16_data['clean_text'].tolist())
# Creating word_cloud with text as argument in .generate() method
word_cloud = WordCloud(width=680,height=656,max_words=150,collocations = True, colormap='tab20c').generate(CORPUS)
# Display the generated Word Cloud
plt.imshow(word_cloud, interpolation='bilinear')
plt.axis("off")
plt.title("Most Frequent Words Used During October 2016",fontsize=18)
plt.show()

```

Output:

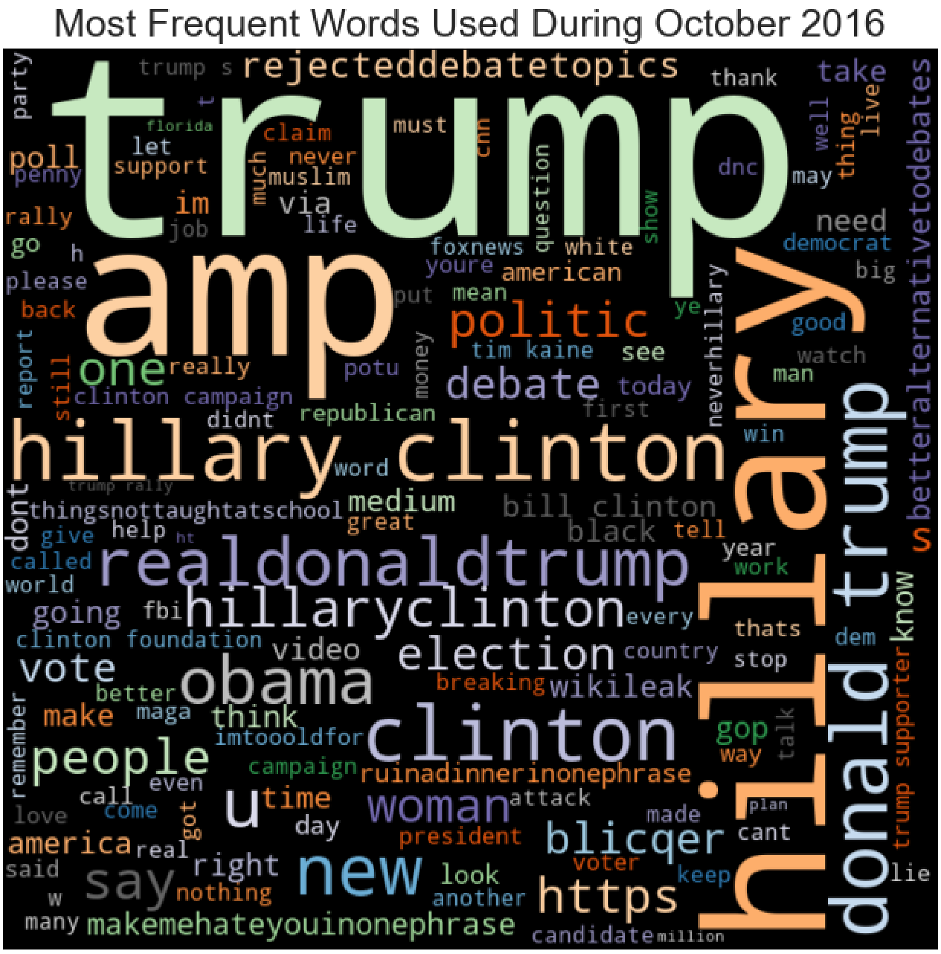
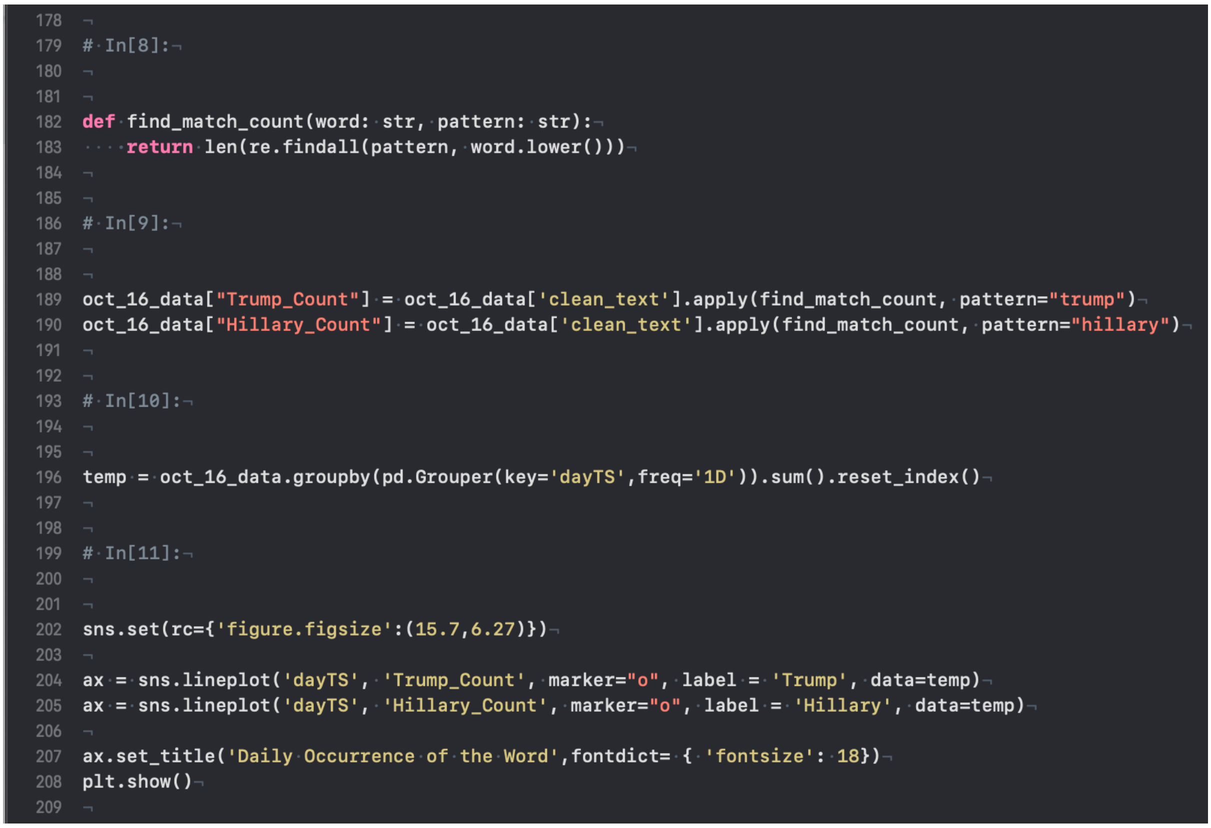


Figure 12 WordCloud of the most tweeted words in October 2016

Time Analysis

We then use this clean text column and input to a function called **find_match_count.** That function returns the number that the words **'trump'** and **'hillary'**, which refer to the names of the 2016 US Presidential nominees, in new columns. We use these newly created count columns to plot a line plot as shown in the below diagram:



```

# In[8]:


def find_match_count(word: str, pattern: str):
    return len(re.findall(pattern, word.lower()))


# In[9]:


oct_16_data["Trump_Count"] = oct_16_data['clean_text'].apply(find_match_count, pattern="trump")
oct_16_data["Hillary_Count"] = oct_16_data['clean_text'].apply(find_match_count, pattern="hillary")


# In[10]:


temp = oct_16_data.groupby(pd.Grouper(key='dayTS',freq='1D')).sum().reset_index()


# In[11]:


sns.set(rc={'figure.figsize':(15.7,6.27)})

ax = sns.lineplot('dayTS', 'Trump_Count', marker="o", label = 'Trump', data=temp)
ax = sns.lineplot('dayTS', 'Hillary_Count', marker="o", label = 'Hillary', data=temp)

ax.set_title('Daily Occurrence of the Word',fontdict= { 'fontsize': 18})
plt.show()

```

Output:

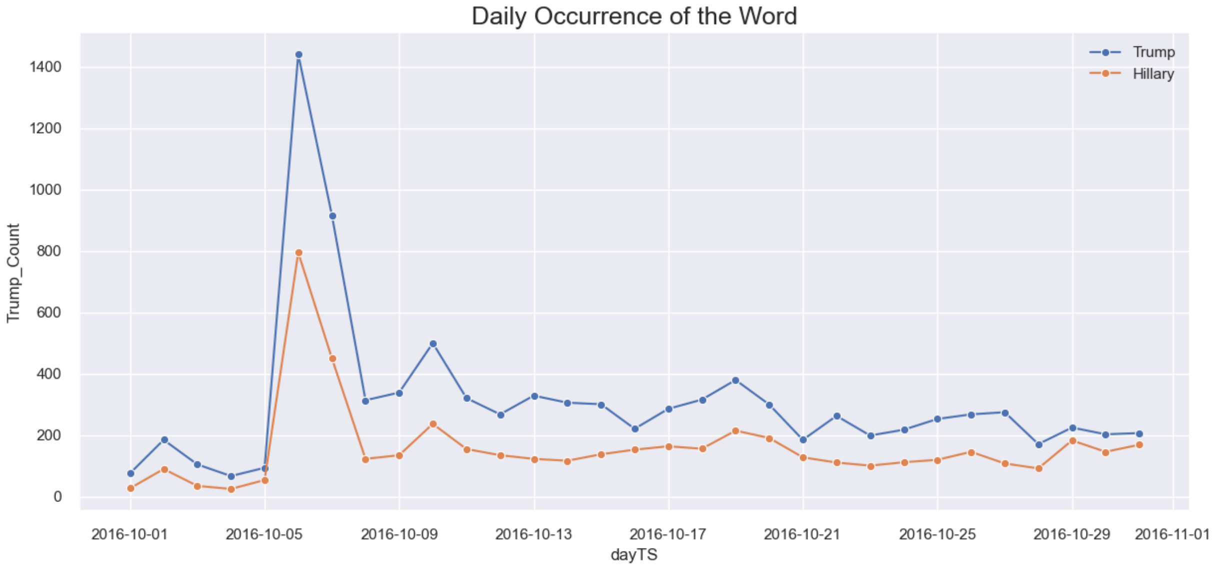


Figure 13 Dual Axis Line Chart showcasing the words' 'trump' and 'hillary' mention count in correlation with time

Remarks:

As the time period of interest in this specific diagram we have chosen October of 2016, which is the final month before the elections since they took place in early November

(November’s 6th). That time space is crucial to be studied in accordance with opinion shaping, popularity and traffic on social media regarding the two campaigns, as the final voting decisions are being made and the public finalizes their support and intention to vote for one candidate or another.

The above chart showcases the amount of times the accounts, that later were exposed as Russian trolls, tweeted in the month October about Hillary or Trump. It is important to highlight that this method focuses on the quantity of the mentions rather than the quality, meaning that it doesn't specify if the name (word) was mentioned in a positive or negative light. Regardless of the approach, higher traffic on social media, independently of if its cause is positive or negative, usually translates in higher percentage of profile engagement, and consequently a bigger platform for the respected candidate, as users visit the candidates profile and interact with their content and get to know their agenda and political views better.

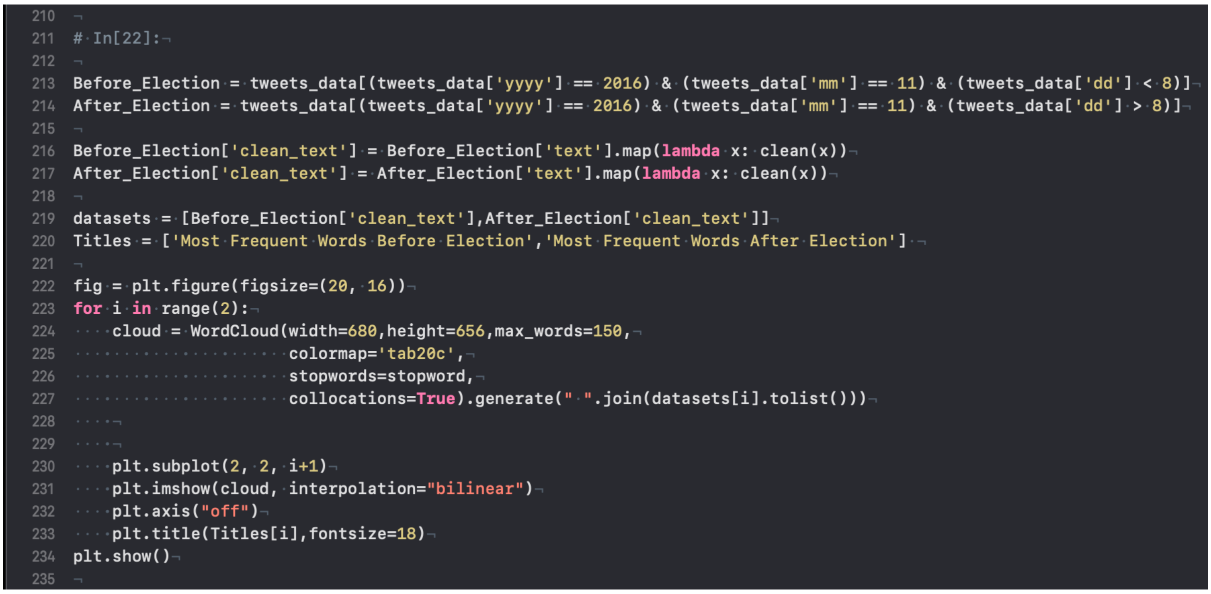
The visualization of the October Time analysis, regarding the name mentions, indicates that at a 100% rate the word ‘trump’ was mentioned in a higher amount of russian troll tweets in comparison with the word ‘hillary’, the mention of the word ‘trump peaked between the 5th and 9th of October 2016, when the word ‘trump’ was mentioned twice as much as the word ‘hillary’, creating higher traffic numbers in that candidate’s account.

Was there a **Before | After** the election ?

In this section, Data is filtered and new data frames are created called **After Election** and **Before Election**,

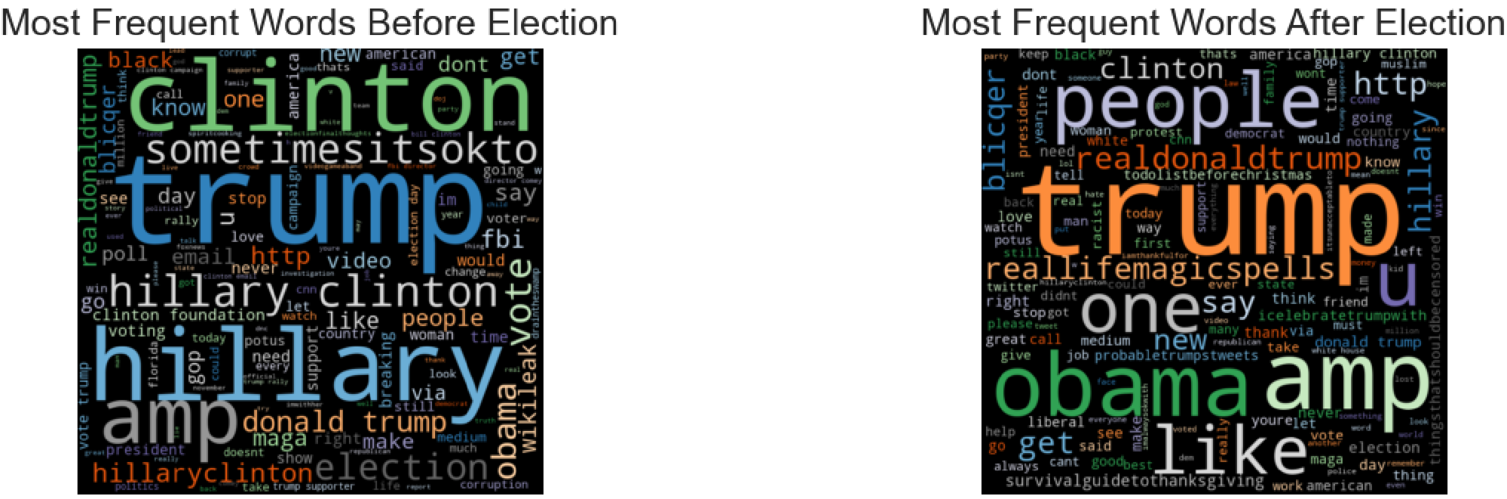
As “After Election” we consider the data filtered for the dates before **8-11-2016** whereas “Before Election”, we consider the data filtered for the Dates after **8-11-2016.**

To plot all the words for both of these Dataframes, we visualize the results using Wordclouds.



```

# In[22]:

Before_Election = tweets_data[(tweets_data['yyyy'] == 2016) & (tweets_data['mm'] == 11) & (tweets_data['dd'] < 8)]
After_Election = tweets_data[(tweets_data['yyyy'] == 2016) & (tweets_data['mm'] == 11) & (tweets_data['dd'] > 8)]

Before_Election['clean_text'] = Before_Election['text'].map(lambda x: clean(x))
After_Election['clean_text'] = After_Election['text'].map(lambda x: clean(x))

datasets = [Before_Election['clean_text'],After_Election['clean_text']]
Titles = ['Most Frequent Words Before Election','Most Frequent Words After Election'] 

fig = plt.figure(figsize=(20, 16))
for i in range(2):
    cloud = WordCloud(width=680,height=656,max_words=150,
                      colormap='tab20c',
                      stopwords=stopword,
                      collocations=True).generate(" ".join(datasets[i].tolist()))
    
    
    plt.subplot(2, 2, i+1)
    plt.imshow(cloud, interpolation="bilinear")
    plt.axis("off")
    plt.title(Titles[i],fontsize=18)
plt.show()

```

Output:



Most Frequent Words Before Election

Most Frequent Words After Election

Figure 14 Comparative WordClouds of the most tweeted words in Before and After the 2016 US Election

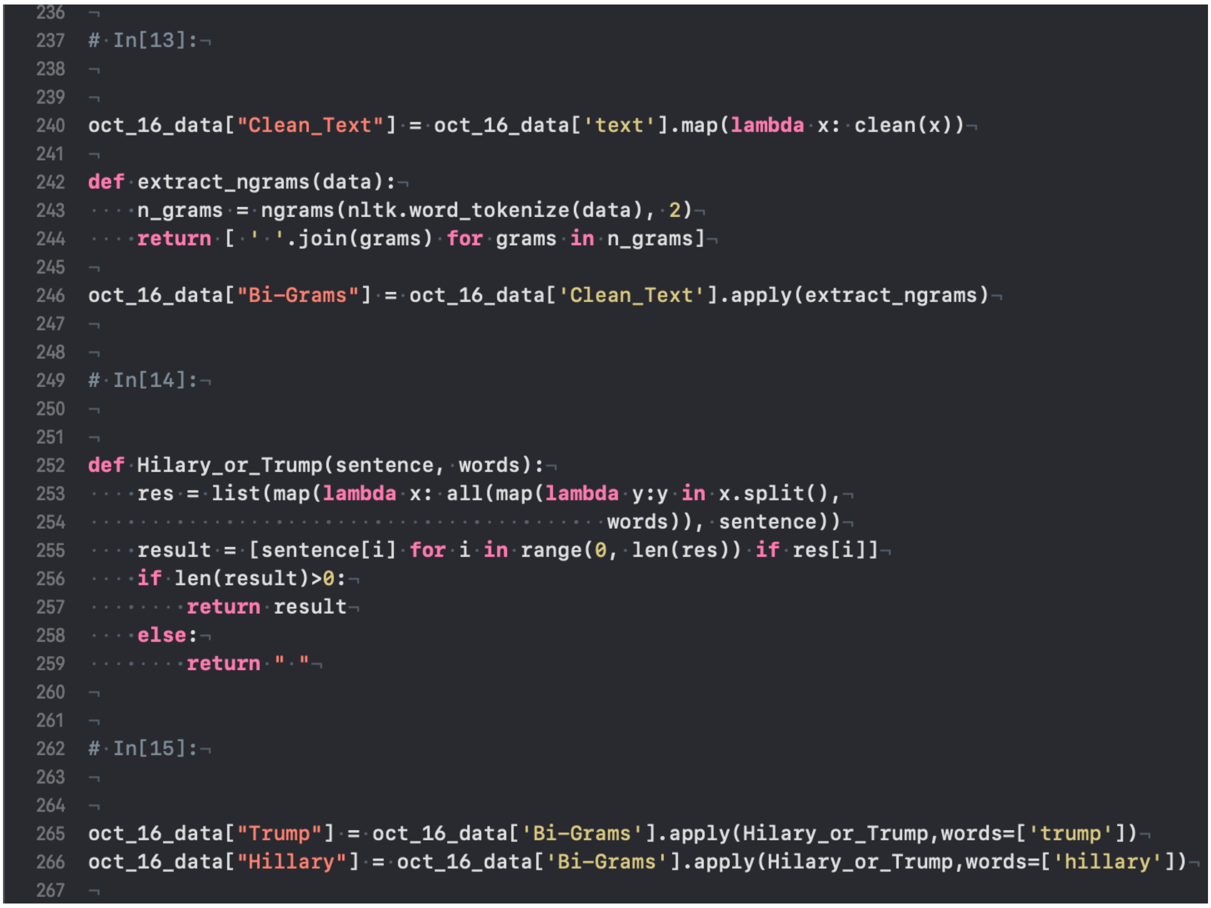
Remarks

By visualizing the most frequent words the Russian trolls tweeted about in October 2016 in the form of word clouds, we attempt to extract the most pertinent parts of the textual data that comes from their tweets. These word clouds also help us compare and contrast pieces of text to find the wording similarities between the two different time periods. Even though word clouds provide generic results we prefer them as they act as initiators to further investigate further the visual information.

The wordcloud before the election, clearly shows that the candidates names are the most tweeted words the trolls tweeted about, but it is obvious that the 'trump' word has been tweeted more than the 'hillary' and the 'clinton' one that follow. Regarding, the after the election word cloud, we can see that a full redistribution of most tweeted words has taken place, as despite the 'trump' word being again the most tweeted one, now Clinton's name has been tweeted about significantly less, and words like 'people' and 'obama', which was the the previous POTUS have been tweeted about significantly more.

2-Grams Sentiment Analysis

Here we try to get Bi-grams from our clean text data
We filter and select the tweet content and create bi-grams, we seperate the bigrams and connect each word following the names of of the Presidential candidates, specifically how they were mostly referred to in the media -> 'trump' or 'hillary' with data from the AFINN lexicon, which gives a numeric sentiment score for each word with positive or negative numbers, indicating the direction of sentiment and then finally plot the scores of words with most occurrences.



```

# In[13]:


oct_16_data["Clean_Text"] = oct_16_data['text'].map(lambda x: clean(x))

def extract_ngrams(data):
    n_grams = ngrams(nltk.word_tokenize(data), 2)
    return [ ' '.join(grams) for grams in n_grams]

oct_16_data["Bi-Grams"] = oct_16_data['Clean_Text'].apply(extract_ngrams)


# In[14]:


def Hilary_or_Trump(sentence, words):
    res = list(map(lambda x: all(map(lambda y:y in x.split(),
                                           words)), sentence))
    result = [sentence[i] for i in range(0, len(res)) if res[i]]
    if len(result)>0:
        return result
    else:
        return " "


# In[15]:


oct_16_data["Trump"] = oct_16_data['Bi-Grams'].apply(Hilary_or_Trump,words=['trump'])
oct_16_data["Hillary"] = oct_16_data['Bi-Grams'].apply(Hilary_or_Trump,words=['hillary'])

```

Output:

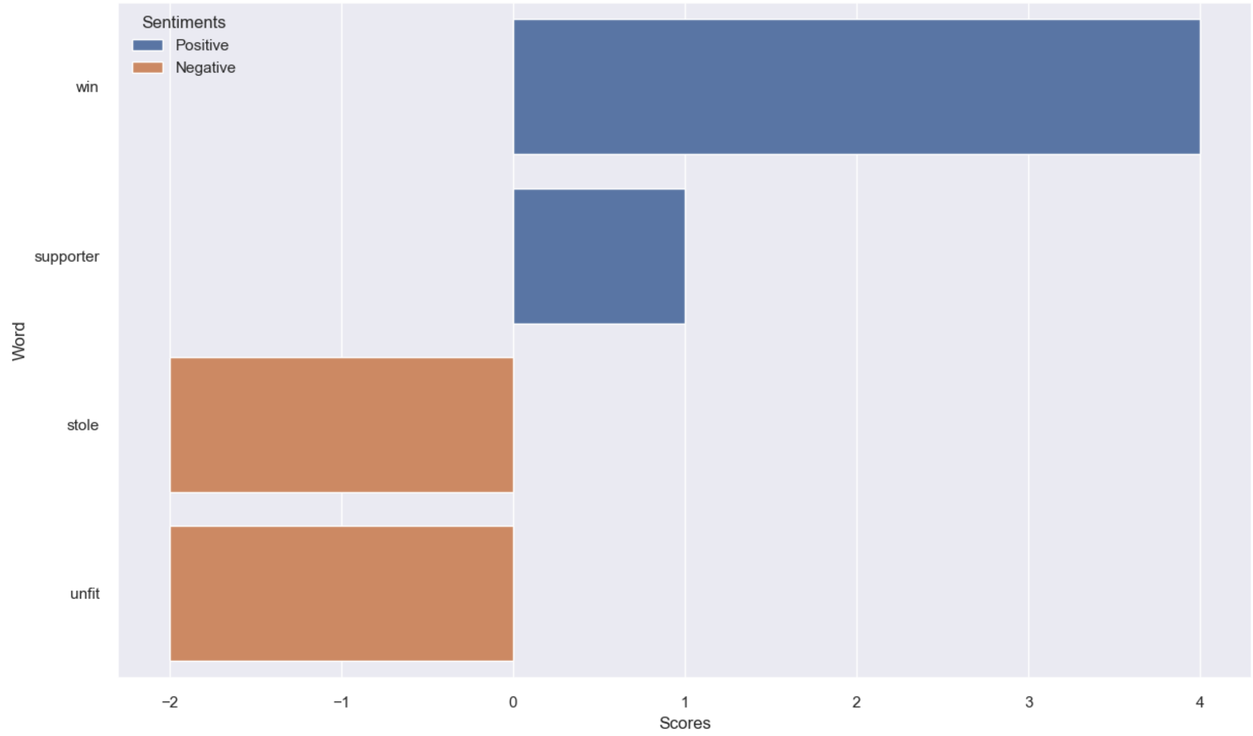


Figure 15 Bar Plot Sentiment Visualization of words following the word 'Hillary'

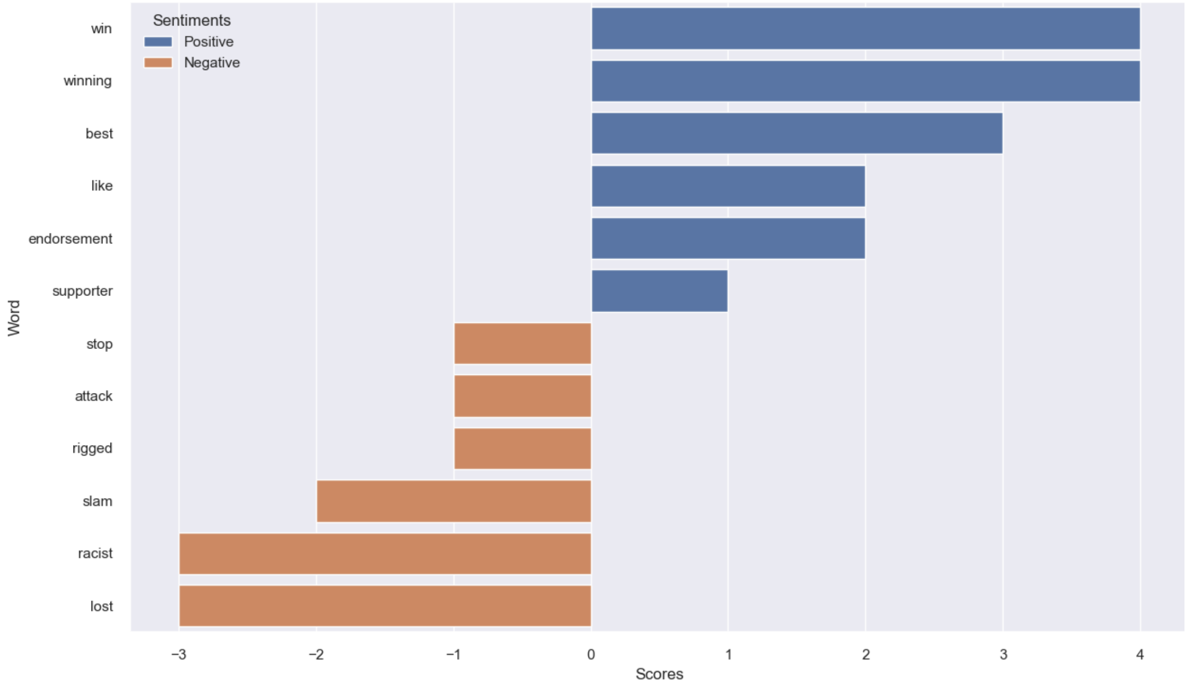


Figure 16 Bar Plot Sentiment Visualization of words following the word 'Trump'

Remarks:

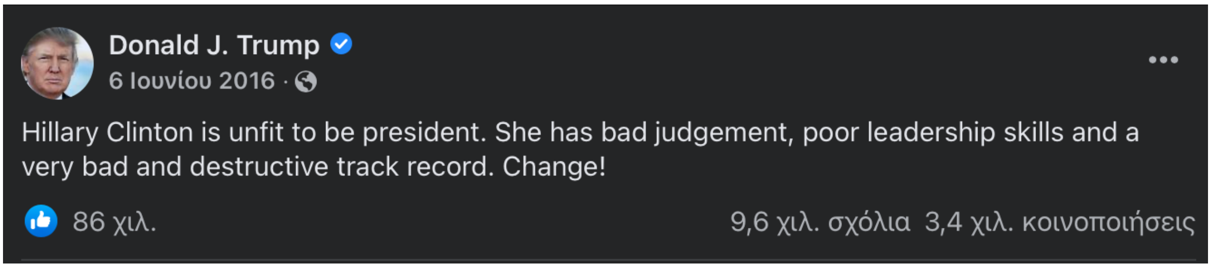
The sentiment analysis of datasets' tweets that include the words 'Trump' or 'Hillary' and specifically the visualization of the sentiment score the words that follow 'hillary' and 'trump' were given by lexicon, show a clear connection not only regarding the intentions of the russian troll accounts but also with the way public opinion was influenced and how the election results finally turned out.

Firstly, what is obvious by just superficially examining the two figures, is that the figure that is referring to the sentiment of the words that follow 'Trump' has **3 times more the amount of most occuring words** than the one that is referring to words that follow 'hillary'. That is justified by Figure 7 that shows that the daily occurrence of the word 'Trump' in the dataset's tweets was constantly higher than the word 'hillary', and in the time period between the 5th and 9th of October of 2016, which was the month prior to the election, the 'trump' word hit an all time high by being mentioned almost double the times than the word 'Hillary'

In Figure 10, that is in regards with the sentiment that characterizes the words following the word 'hillary' we can witness that the most occuring words that are associated with a positive sentiment are 'wins' and 'supporter' with a lexicon sentiment score of 4 and 1, respectively and the most occuring words that are associated with a negative sentiment are the words 'stole' and 'unfit' both with a lexicon sentiment score of -2.

In Figure 11, that is in regards with the sentiment that characterizes the words following the word 'Trump' we can witness that the most occuring words that are associated with a positive sentiment are 'win', 'winning', 'best', 'like', 'endorsement' and 'supporter' with a lexicon sentiment score of 4, 4, 3, 2, 2 and 1, respectively and the most occuring words that are associated with a negative sentiment are the words 'stop', 'attack', 'rigged', 'slam', 'racist', 'lost' with a lexicon sentiment score of -1, -1, -1, -2 , -3 and -3, respectively.

It is of crucial importance to mention, that despite the fact that lexicon indicates the words 'stop', 'attack', 'rigged' and 'slam' as negative ones and consequently categorizes them as negative tweets against Trump, that is not the case, as the words mentioned above have all been mentioned by Donald Trump multiple times during his speeches, rallies and even during the 3 Presidential Debates. Taking under consideration, that his supporters tend to imitate and emulate his way of expressing his opinions and that russian trolls that produced his tweets wanted to mimic real pro-Trump accounts, it is indicated that those negative words were not against Trump but in reality was perpetuating his aggressive narrative and increasing the interaction of users with his ideas. On the other side, the words that followed Hillary's name and were marked as negative, deeply connected with Trumps' emotion fueling way of expression, as he called Hillary several times both on social media and on public speeches unfit to be a President of the United States. That word was used by the Russian trolls and perpetuated the narrative that Donald Trump initiated.



Donald J. Trump
6 Ιουνίου 2016 ·

Hillary Clinton is unfit to be president. She has bad judgement, poor leadership skills and a very bad and destructive track record. Change!

86 χιλ.   9,6 χιλ. σχόλια 3,4 χιλ. κοινοποιήσεις

Figure 17 Facebook post of Donald Trump that calles Hillary Clinton 'unfit' for President of the United States

4.2 AFINN

AFINN[48] is a list of English terms manually rated for valence with an integer between negative five (-5) and positive five (+5) The original lexicon used to contain multi word phrases that are excluded here.

Afinn is the simplest - yet one of the most popular - lexicons used for sentiment analysis developed by *Finn Årup Nielsen*.

It contains 3300+ words with a polarity score associated with each word. In python, there is an in-built function for this lexicon.

[48] Finn Årup Nielsen A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages 718 in CEUR Workshop Proceedings* 93-98. 2011 May.http://arxiv.org/abs/1103.2903.

4.3 Sentiment Analysis NLTK

Natural Language Processing (NLP) is part of computer science and artificial intelligence that deals with human languages.

Natural language processing is a model for studying language ability and language application. A computer (algorithm) framework is built to implement such a language model, and it is perfected, evaluated, and finally used to design various practical systems. Its branches are Automatic Speech Recognition (ASR), Named entity recognition (NER), Optical character recognition (OCR), Sentiment analysis and so on.

NLP is a component of text mining that conducts a specific kind of linguistic analysis that helps the computer "read" text.

At the same time, Natural Language Processing uses a different methodology to decipher the ambiguities in human languages. This methodology includes methods such as, automatic summarization, part-of-speech tagging, disambiguation, chunking and disambiguation and natural language understanding and recognition.

Sentiment analysis, sometimes known as **opinion mining or emotion AI**, **refers** to the use of natural language processing, text analysis and computational linguistics to identify, extract, quantify, and study in a systematic way subjective preferences and affective states.

Sentiment analysis aims to determine the attitude of a writer, or of another subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event.

4.4 Lemmatization

Lemmatization is the process of converting a word to its base form. Contrary to another technique called stemming, in which the last few characters of a word are removed, in lemmatization the context is taken under consideration and the word is converted to its meaningful base form. This function allows lemmatization to avoid providing incorrect meanings and spelling errors.

For instance, lemmatization would correctly identify the base form of 'caring' to 'care', contrary to stemming which would just cut off the '-ing' part and convert it to car.

Lemmatization can be implemented in python by using Wordnet Lemmatizer, Spacy Lemmatizer, TextBlob and Stanford CoreNLP.

4.5 Bigrams

“Language modeling is the way of determining the probability of any sequence of words. Language modeling is used in a wide variety of applications such as Speech Recognition, Spam filtering, etc. In fact, language modeling is the key aim behind the implementation of many state-of-the-art Natural Language Processing models.”[49]

N-gram

N gram could be defined as the contiguous sequence of n items from a given sample of text or speech. The aforementioned items could be letters, words or base pairs according to the application. N grams are typically collected from a long text dataset (text or speech corpus)

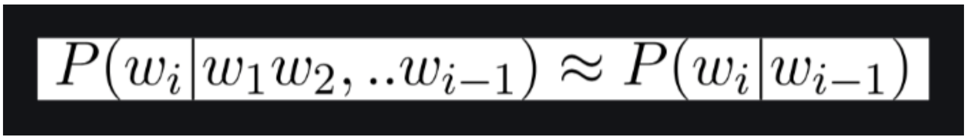
N-gram Language Model:

An N-gram language model predicts the probability of a given N-gram within any sequence of words in the language. **A good N-gram model can predict the next word in the sentence** i.e the value of p(w|h)

Examples of N-gram as unigram are : “This”, “article”, “is”, “on”, “NLP”) and as bi-gram: ‘This article’, ‘article is’, ‘is on’,’on NLP’.

Now, we will establish a relation on how to find the next word in the sentence using . We need to calculate p(w|h), where is the candidate for the next word. If we would want to calculate the probability of the last word being ‘NLP’ having the previous words we would, after simplifying several equation end up having

Regarding the Bigram:



$$P(w_i|w_1w_2,..w_{i-1}) \approx P(w_i|w_{i-1})$$

[49] https://www.geeksforgeeks.org/n-gram-language-modelling-with-nltk/

Chapter 5: Conclusions & Discussion

5.1 Conclusions

According to Galeotti, "Russia's attempts to distract, divide, and demoralize have been called a form of political war". This analysis has given insight into the methods the IRA used to attack the United States by undermining genuine cultural and political discourse. One former employee of the IRA described the feeling of working there as though "you were in some kind of factory that turned lying, telling untruths, into an industrial assembly line"[50] The systematic nature of the tweets we have analyzed suggests that this feeling was correct.

The IRA engaged in what is not simply political cyberwarfare, but industrialized political cyberwarfare.

In terms of trolling content, "enormous heterogeneity in theme and approach across IRA accounts". For instance, some of the tweets were targeted at right-wing followers and others to sow discord on the left. There is also substantial variation over time in the creation of the trolling accounts and trolling intensity.

We attempted to provide an elaborate combination of Knowledge Graph and Natural Language Processing methodologies, such as modeling, data cleaning, visualization, time analysis, content analysis and sentiment analysis, to further understand the semantic relationships among entities in trolls. The Visualization approach has been conducted to further understand the events or statements expressed in the relationships among different entities in troll tweet sets from the 2016 Presidential Election. The trolls targeted one candidate, Hillary Clinton, and her family, by repeatedly accusing her of the "email-gate scandal" and of misuse of the Clinton foundation, etc. The negative sentiments in troll tweets that were referring to Hillary Clinton was the highest, which shows that Russian Trolls had used many more negative terms portraying Clinton than Trump.

[50] Linvill, D. & Warren, Patrick. (2020). Troll Factories: Manufacturing Specialized Disinformation on Twitter. Political Communication. 37. 1-21. 10.1080/10584609.2020.1718257.

5.2 Discussion & Future Work

The IRA efforts in the time period we studied can be characterized as systematic. Their system was industrial – mass produced from a collection of interchangeable parts. It is obvious from our analysis that the IRA focused on divergent, often contrary agenda in their disinformation campaigns, engaging with opposing, ideologically engaged networks and even continuing posting even after the end of the election. This gives credibility to the narrative that one effort the IRA was engaged in was to divide the United States of America along partisan lines by weaponizing multiple ideologies against each other.

Shedding light on how governments, government-affiliated and politically motivated organizations work to influence the global power distribution is essential, and the IRA operation is an important example of digital age interference, that is going to be referred to as the beginning of information and data warfare in cyberspace.

For this reason, future research will need to examine IRA efforts further, as well as the efforts of other initiators of state-affiliated disinformation.

None-the-less, future research should endeavor to explore methods of reliably identifying valid sets of disinformation produced on social media platforms. Any approach to doing so would likely have additional limitations, but understanding this important element of our political discourse cannot remain reliant on the content which for profit media platforms do or do not choose to share publicly. Future research should also aim to better understand any potential effects of state-sponsored disinformation and other forms of public agenda building. Such questions could not begin to be answered with the data analyzed in this study, however.

Our approach to trolling attempts to isolate variation in disinformation coming from abroad, in comparison with other domestic interferences that can disturb election processes all over the globe. In accordance with the Senate Intelligence Committee’s 2019 (bipartisan) report that Russian trolling was attempting to help the Trump campaign win the US presidential elections of 2016, our approach partly confirms that. As it readily generalizes the analysis of additional outcomes that may be of interest to researchers and are believed to impact elections, such as time-series variation in political campaign donations, US street protests, etc. Our analysis can be treated as initial “proof of concept” for future analyses that emphasizes:

1. Analyzing non-English tweets of this data, in various contexts, to better understand how the IRA’s tactics adapt over time
2. “Drivers” of Russian trolling activity - neglected by the current literature.
3. Trolling’s causal effect on outcomes that the precise path and downstream impacts cannot be traced via social media platforms.

As part of a global society, it is crucial that scientists across disciplines study and investigate such threats to the common good. With continual research, we will be able to better anticipate and identify future challenges in the democratic process, as well as international influence operations. Researchers around the globe are continuing to make discoveries about the nature of the 2016 U.S. presidential election interference, creating opportunities for additional study, reflection, and planning.

This thesis paper can be considered a call on members of the international community to contribute in unmasking those who attempt to undermine or harm the most basic human rights, those of self determination, freedom, and fairness.

5.3 Academic Acknowledgement

The author acknowledges all individuals whose investigations have helped to discover and identify sources of interference in the U.S. election process, including (but by no means limited to) Special Counsel Robert S. Mueller III, members of the U.S. Department of Homeland Security, Darren Linvill, Patrick Warren, the international Intelligence Community, and the United States House Permanent Select Committee on Intelligence.

Chapter 6: References

6.1 Bibliography

1. Thebusinessresearchcompany.com. 2022. *Social Media Market Size 2022 And Growth Analysis*. [online] Available at: <https://www.thebusinessresearchcompany.com/report/social-media-global-market-report> [Accessed 25 September 2022].
2. Insights, F., 2022. *With 13.4% CAGR, Big Data Analytics Market Size Worth USD 655.53 Billion by 2029*. [online] GlobeNewswire

News Room. Available at: <https://www.globenewswire.com/en/news-release/2022/07/21/2483358/0/en/With-13-4-CAGR-Big-Data-Analytics-Market-Size-Worth-USD-655-53-Billion-by-2029.html> [Accessed 25 September 2022].

3. Hayati, P., Potdar, V., Talevski, A., Firoozeh, N., Sarenche, S. and Yeganeh, E., 2022. *Definition of spam 2.0: New spamming boom.*
4. Hayati, P., Potdar, V., Talevski, A., Firoozeh, N., Sarenche, S. and Yeganeh, E., 2022. *Definition of spam 2.0: New spamming boom.*
5. T. O'Reilly, "What Is Web 2.0," in http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web- 20.html: O'Reilly Network, 2005. [Accessed 25 September 2022].
6. P. Hayati, K. Chai, V. Potdar, and A. Talevski, "HoneySpam 2.0: Profiling Web Spambot Behaviour," in *12th International Conference on Principles of Practise in Multi-Agent Systems*, Nagoya, Japan, 2009, pp. 335- 344. [Accessed 25 September 2022].
7. "How to Spot a Fake PayPal Email," [Online]. Available: https://www.secureworldexpo.com/industry-news/how-to-spot-a-fake- paypal-email. [Accessed 25 September 2022].
8. https://www.dhl.com/gr-en/home/footer/fraud-awareness.html [Accessed 25 September 2022].
9. Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. (2016). "The rise of social bots." Commun. ACM 59, 7 (July 2016), 96–104 [Accessed 25 September 2022].
10. Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Alessandro Flammini, and Filippo Menczer (2017) "The spread of fake news by social bots"[Accessed 25 September 2022].
11. Bessi, Alessandro and Ferrara, Emilio (2016), "Social Bots Distort the 2016 US Presidential Election Online Discussion" First Monday, Volume 21, Number 11 -- 7 November 2016 [Accessed 25 September 2022].
12. Himelein-Wachowiak, M., Giorgi, S., Devoto, A., Rahman, M., Ungar, L., Schwartz, H. A., Epstein, D. H., Leggio, L., & Curtis, B. (2021), "Bots and Misinformation Spread on Social Media: Implications for COVID-19. Journal of medical Internet research" [Accessed 25 September 2022].
13. Donath Judith S. Identity and deception in the virtual community. *Communities in Cyberspace* 1999 [Accessed 25 September 2022].
14. Schwartz Mattathias. *NY Times Magazine.* 2008. The trolls among us.
15. Hardaker Claire. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *J Politeness Res* 2010 [Accessed 25 September 2022].

16. Gray, K. L. (2017). Gaming out online: Black lesbian identity development and community building in Xbox Live. Journal of Lesbian Studies, 22(3), 282–296. https://doi.org/10.1080/10894160.2018.1384293 [Accessed 25 September 2022].

17. Gray, K. L., Buyukozturk, B., Hill, Z. G. (2017). Blurring the boundaries: Using Gamergate to examine "real" and symbolic violence against women in contemporary gaming culture. Sociology Compass, 11(3), Article e12458. [Accessed 25 September 2022].

18. https://www.cmu.edu/news/stories/archives/2020/may/twitter-bot-campaign.html [Accessed 25 September 2022]

19. Aro, J. (2016), The Cyberspace War: Propaganda and Trolling as Warfare.

Tools, European View, (15), 121-132. [Accessed 25 September 2022].

20. Duskaeva, L.R., Konyaeva, L.R. (2016), Trolling in Russian Media, Journal of Organizational Culture, Communications and Conflict, (4), 58–67. [Accessed 25 September 2022].

21. Karpan, A. (2018),Troll Factories: Russia's Web Brigades, Greenhaven Publishing, New York. [Accessed 25 September 2022].

22. Bernal, P. (2018), The Internet, Warts and All: Free Speech, Privacy and Truth, Cambridge University Press, Cambridge. [Accessed 25 September 2022].

23. Lehto, M., Neittaanmäki, P. (Eds.) (2018), Cyber Security: Power and Technology, Springer, Cham. [Accessed 25 September 2022].

24. https://intelligence.house.gov/social-media-content/ [Accessed 25 September 2022].

25. "Platform manipulation and spam policy," https://help.twitter.com/en/rules-and-policies/platform-manipulation [Accessed 25 September 2022].

26. https://ec.europa.eu/info/law/law-topic/data-protection/eu-data-protection-rules_en#gdpr-the-fabric-of-a-success-story

27. Beinke, Thies & Schamann, Annabell & Freitag, Michael & Feldmann, Klaas & Brandt, Matthias. (2017). Text-Mining and Gamification for the Qualification of Service Technicians in the Maintenance Industry of Offshore Wind Energy. International Journal of e-Navigation and Maritime Economy. 6. 44-52. 10.1016/j.enavi.2017.05.006. [Accessed 25 September 2022].

28. Rüdiger, Matthias & Antons, David & Salge, Oliver. (2017). From Text to Data: On The Role and Effect of Text Pre-Processing in Text Mining Research. Academy of Management Proceedings. 2017. 16353. 10.5465/AMBPP.2017.16353abstract.

29. https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

30. Antons, David & Grünwald, Eduard & Cichy, Patrick & Salge, Oliver. (2020). The application of text mining methods in

innovation research: current state, evolution patterns, and development priorities. R&D Management. 50. 10.1111/radm.12408.

31. Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1 - 30.Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1 - 30.
32. Antons, David & Breidbach, Christoph & Joshi, Amol & Salge, Oliver. (2021). Computational Literature Reviews: Method, Algorithms, and Roadmap. Organizational Research Methods. 109442812199123. 10.1177/1094428121991230.
33. Gamache, Daniel & Mcnamara, Gerry & Mannor, Michael & Johnson, Russell. (2014). Motivated to Acquire? The Impact of CEO Regulatory Focus on Firm Acquisitions. The Academy of Management Journal. 58. 10.5465/amj.2013.0377.
34. Pennebaker, James & Boyd, Ryan & Jordan, Kayla & Blackburn, Kate. (2015). The Development and Psychometric Properties of LIWC2015. 10.15781/T29G6Z.
35. Darren L. Linvill & Patrick L. Warren (2020) Troll Factories: Manufacturing Specialized Disinformation on Twitter,Political Communication, 37:4, 447-467, DOI: 10.1080/10584609.2020.1718257
36. Ahmer Arif, Leo Graiden Stewart, and Kate Starbird. 2018. Acting the Part: Examining Information Operations Within #BlackLivesMatter Discourse. Proc. ACM Hum.-Comput. Interact. 2, CSCW, Article 20 (November 2018), 27 pages. https://doi.org/10.1145/3274289
37. Marwick, A. E., & boyd, danah. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, *13*(1), 114–133. https://doi.org/10.1177/1461444810365313
38. https://www.nbcnews.com/tech/social-media/now-available-more-200-000-deleted-russian-troll-tweets-n844731
39. Linvill, D. & Warren, Patrick. (2020). Troll Factories: Manufacturing Specialized Disinformation on Twitter. Political Communication. 37. 1-21. 10.1080/10584609.2020.1718257.

6.2 Further Reading

1. USHPSCI, "Exposing Russia's Effort to Sow Discord Online: The Internet Research Agency and Advertise- ments," *Exposing Russia's Effort to Sow Discord Online: The Internet Research Agency and Advertisements | U.S. House of Representatives*, 2018. [Online]. Available: https://democrats-intelligence.house.gov/social-media-content/. [Accessed: 25-Sept-2022].
2. USHPSCI, "Social Media Advertisements," 2018. [Online]. Available:https://democrats-intelligence.house.gov/social-media-c

ontent/social-media- advertisements.htm. [Accessed: 25-Sept-2022].

3. O. Roeder, "Why We're Sharing 3 Million Russian Troll Tweets," *FiveThirtyEight*, 31-Jul-2018. .
4. M. Staton, "Faculty measure impact of underreported ac- tivity of political Twitter trolls," *Newsstand | Clemson University News and Stories, South Carolina*, 2018. [Online]. Available: http://newsstand.clemson.edu/faculty-measure-impact-of-underreported-activity-of-political- twitter-trolls/. [Accessed: 25-Sept-2022].
5. D. L. Linvill and P. L. Warren, "Troll Factories: The In- ternet Research Agency and State-Sponsored Agenda Building," 2018.
6. R. L. Boyd, *MEH: Meaning Extraction Helper [Software]*. 2018.
7. R Core Team, *R: A Language and Environment for Statis- tical Computing*. Vienna, Austria: R Foundation for Statis- tical Computing, 2018.
8. R. L. Boyd, S. R. Wilson, J. W. Pennebaker, M. Kosinski, D. J. Stillwell, and R. Mihalcea, "Values in Words: Using Language to Evaluate and Understand Personal Values," in *Proceedings of the Ninth International AAAI Confer- ence on Web and Social Media*, 2015, pp. 31–40.
9. C. K. Chung and J. W. Pennebaker, "Revealing Dimen- sions of Thinking in Open-Ended Self-Descriptions: An Automated Meaning Extraction Method for Natural Lan- guage," *Journal of Research in Personality*, vol. 42, no. 1, pp. 96–132, Feb. 2008.
10. A.Kramer and C. Chung, "Dimensions of Self-Expression in Facebook Status Updates," in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011, pp. 169–176.
11. N. Ramirez-esparza, C. K. Chung, E. Kacewicz, and J. W. Pennebaker, "The psychology of word use in depression forums in English and in Spanish: Testing two text analyt- ic approaches," in *In Proc. ICWSM 2008*, 2008.
12. I. Jolliffe, "Principal Component Analysis," in *Interna- tional Encyclopedia of Statistical Science*, M. Lovric, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1094–1096.
13. J. Shlens, "A Tutorial on Principal Component Analysis," *arXiv:1404.1100 [cs, stat]*, Apr. 2014.
14. S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1, pp. 37–52, Aug. 1987.
15. R. L. Boyd, "Psychological text analysis in the digital humanities," in *Data analytics in the digital humanities*, S. Hai-Jew, Ed. New York: Springer International Publishing, 2017, pp. 161–189.
16. C. Wilke, *Geoms to make ridgeline plots with ggplot2. Contribute to clauswilke/ggridges development by creating an account on GitHub*. 2018.

17. A. Ng, "This was the most viewed Facebook ad bought by Russian trolls," *CNET*, 2018. [Online]. Available: https://www.cnet.com/news/this-was-the-most-viewed-facebook-ad-bought-by-russian-trolls/. [Accessed: 25-Sept-2022].

18. FiveThirtyEight, "3 million Russian troll tweets," 2018. [Online]. Available: https://github.com/fivethirtyeight/russian-troll-tweets. [Accessed: 25-Sept-2022].

19. G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh, and L. Chanona-Herna´ndez, "Syntactic N-grams as machine learning features for natural language processing," *Expert Systems with Applications*, vol. 41, no. 3, pp. 853–860, Feb. 2014.

20. R. L. Boyd, "Mental profile mapping: A psychological single-candidate authorship attribution method," *PLOS ONE*, vol. 13, no. 7, p. e0200588, Jul. 2018.

21. M. Coulthard, A. Johnson, D. Wright, A. Johnson, and D. Wright, *An Introduction to Forensic Linguistics : Lan- guage in Evidence*. Routledge, 2016.

22. M. Koppel, J. Schler, and E. Bonchek-Dokow, "Measuring differentiability: Unmasking pseudonymous authors," *Journal of Machine Learning Research*, vol. 8, no. 2007, pp. 1261–1276, 2007.

23. M. Koppel and Y. Winter, "Determining if two documents are written by the same author," *Journal of the American Society for Information Science and Technology*, vol. 65, no. 1, pp. 178–187, 2014.

24. C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, *The Stanford CoreNLP Natural Lan- guage Processing Toolkit*. 2014.

25. L. Derczynski, A. Ritter, S. Clark, and K. Bontcheva, "Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data," in *Proceedings of the Interna- tional Conference on Recent Advances in Natural Lan- guage Processing*, 2013.

26. J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Black- burn, "The development and psychometric properties of LIWC2015," Austin, TX, 2015.

27. S. R. Narum, "Beyond Bonferroni: Less conservative analyses for conservation genetics," *Conserv Genet*, vol. 7, no. 5, pp. 783–787, Oct. 2006.

28. C. Gabrielatos and A. Marchi, "Keyness: Appropriate metrics and practical issues," in *Proceedings of the 2012 International Conference on Corpus-assisted Discourse Studies*, University of Bologna, Italy, 2012.

29. T. Ionin, M. L. Zubizarreta, and S. B. Maldonado, "Sources of linguistic knowledge in the second language acquisition of English articles," *Lingua*, vol. 118, no. 4, pp. 554–576, 2008.

30. M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, "Lying words: Predicting deception from lin- guistic

styles,” *Personality and Social Psychology Bulletin*, vol. 29, no. 5, pp. 665–675, 2003.

31. A. Field, D. Kliger, S. Wintner, J. Pan, D. Jurafsky, and Y. Tsvetkov, “Framing and Agenda-setting in Russian News: a Computational Analysis of Intricate Political Strategies,” *arXiv:1808.09386 [cs]*, Aug. 2018.