Vector-Structured Data, Tweet Analysis

JHARNA KUMARI

Abstract- This report delves into the comprehensive analysis on text data using vector embeddings performed on Twitter (now called X) dataset related to the Internet Research Agency (IRA,) popularly known as Russian Troll Factory. In the lead-up to the 2016 US presidential election, the agency allegedly intentionally tried to sow political discontent in the United States with inflammatory social media content. Twitter provided the dataset containing details about thousands of profiles associated with IRA. The main objective of this project is to analyse the dataset provided and show the presence of bias be it political or gender or any other. The dataset employed for the study involves a diverse range of tweet metrics, involving content, author id, date posted, updates, etc. This is used for understanding the underlying structure of the dataset and identifying interesting patterns or relationships that will further increase scope for future research. To achieve this, the report begins by data importation where Twitter data is imported with all the tweets produced by IRA related twitter accounts since 2012 in a public GitHub repository issued by Clemson University researchers Darren Linvill, an associate professor of communication, and Patrick Warren, an associate professor of economics. Furthermore, the report explores data sanitisation and tokenization which forms the base for vector structured data. After word embedding is trained using Python libraries, a subset of the vector data is clustered. Bias is observed by comparison and visualisations. The study concludes by discussing the implications and limitations of the findings and potential avenues for future research.

ACM Reference Format:
Jharna Kumari. 2024. Vector-Structured Data, Tweet Analysis. 1, 1 (July 2024), 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The 2016 United States presidential election was the 58th presidential election that was held on Tuesday, November 8, 2016. In one of the biggest political turmoils in American history, Republican candidates Donald Trump triumphed over Democratic candidate Hillary Clinton, former secretary of state and first lady. Clinton led in almost every nationwide and swing-state poll, with some predictive models giving Clinton over a 90 percent chance of winning. On Election Day, Trump over-performed his polls, winning several key swing states, while losing the popular vote by 2.87 million votes[1]. According to the U.S. intelligence community, the Russian Government used espionage to indirectly interfere in this elections by sabotaging the campaign of democratic candidate Hillary Clinton, boosting the presidential campaign of Donald Trump. Twitter, a big social media platform, released the details of thousands of accounts affiliated with IRA, originating in Russia, containing 10 million tweets since 2009 - 9 million directly linked with IRA. The Russian troll factory is responsible for flooding Twitter with troll content with anti-Clinton propaganda as well infiltrated and further polarised already polarized internet communities. The massive data dump reveals how trolls disrupt and destabilize local and global politics. It has become popularly very common to use trolls to influence politics, healthcare, society at large through social media. It is still unclear as to what a troll actually implies. Kumar et al. (Kumar, 2014) use the term "trolling" when a user posts and spreads information that is deceptive, inaccurate, or outright rude. These trolls use anonymity as a disguise and make use of availability of the internet to provoke, harass and spread hate in the world. They derive satisfaction through their hate speech by causing emotional distress in others. One of the ways to identify trolls or understand the behavior of certain users online is Natural language processing (NLP), defined as the ability of a computer program to understand human language as its spoken as well as written. It can be considered as a subfield of artificial intelligence used to comprehend natural language. NLP is often used for speech recognition, topic detection, text classification, bias detection and much more. NLP uses machine

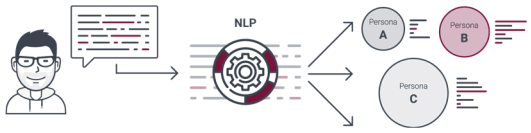

Fig. 1. NLP Operation

learning algorithms to analyse large amounts of unstructured data and extracts meaningful insights. It can recognise patterns and make inferences based on those patterns. NLP works as -

- A sentence or group of sentences are fed into the Natural Language Processing system.
- The system breaks down the sentences into smaller parts of words, called tokens.
- These tokens are then subjected to certain algorithms to find hidden patterns.

In this report we analyse a large dataset often referred to as big data for gaining insights that may be diverse and valuable. The need to understand preferences and behavior can reveal patterns that can be useful or avoid disasters. One of the crucial parts of data analytic, effective data collection provides a foundation for deriving hidden information from the existing framework which later is used to predict or classify unknown future trends. In science, medicine, climatic, production, and other fields, typically, researchers create and implement measures to collect specific sets of data as part of their data collection processes. Primarily there are two types of data

Author's address: Jharna Kumari, jkumari@uvic.c.

ACM XXXX-XXXX/2024/7-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

– quantitative and qualitative data. Quantitative data is any data that can be counted or measured whereas qualitative data is descriptive mostly data that can be observed but not measured. In this report, we would be focusing on qualitative data which is text data. The text data is then vectorized through word embeddings. Word embeddings are a powerful machine-learning framework that represents each English word by a vector. The geometric relationship between these vectors captures meaningful semantic relationships between the corresponding words[2].

2 DATASET

The dataset used for this project was revealed by Twitter to the US government claiming that the accounts and tweets included were related with the "Internet Research Agency". The IRA is a company paid by the Russian government to sow disinformation (Wikipedia, 2019). The data is publicly available and contains over 3 million tweets from thousands of twitter handles. It includes author' information, date of tweet, content of tweet, redirection links, number of followers and following, a flag if the tweet was a retweet. The dataset is called Russian Troll dataset. Another interesting information is the column account_type that shows whether the account is leftist or right. This dataset has been a valuable resource for researchers and analysts studying online influence through hate speech. It offers insights into how international organisations can use social media to spread disinformation, manipulate public discourse, and potentially influence political outcomes of another country. The disclosure of this Russian troll factory dataset was part of Twitter's effort to address concerns regarding foreign interference and the spread of fake news via its platform.

3 METHODOLOGY

The following methodology is used to find the political biases in the given Twitter data:

(1) Data Loading and Pre-processing
- *Load Dataset* : The first step is to load the CSV files. In total, there are 13 CSV files, sourced for the Internet Research Agency, containing tweets from alleged Russian troll accounts leading up to the 2016 presidential election.
- *Text Preprocessing* : After loading the CSV files, the next step is to preprocess the text of the CSV files. The preprocessing step will tokenize, convert to lowercase, remove stop words, and lemmatize the data before treating it to the model.

(2) Word Embedding Training
- *Train Word2Vec Model* : The preprocessed text data is trained on the word2vec model, where each word in the dataset is represented as a unique vector in a high-dimensional space. the model will learn to map words with similar contexts to the nearby vector space. After the training of the model, the model is saved.

(3) Loading pre-trained model
- *Spectral clustering on a pre-trained model* : For the comparison between the model built by us and the pre-trained model, the pre-trained model is loaded and spectral clustering is applied to a specific list of words related to politics chosen from the dataset.

(4) Spectral Clustering
- *Subset selection* : A subset of words for clustering was manually selected based on the relevance to the analysis. The list of words chosen for analysis is **'Democracy, Republic, Monarchy, Autocracy, Oligarchy, Dictatorship, Totalitarianism, Anarchy, Government, Constitution, Parliament, Congress, Senate, Legislature, Judiciary, Executive, President, Prime Minister, Governor, Mayor, Minister, Election, Campaign, Vote, Ballot, Candidate, Party, Platform, Debate, Policy, Law, Bill, Amendment, Referendum, Plebiscite, Suffrage, Citizenship, Nationalism, Federalism, State, Municipality, Bureaucracy, Regulation, Taxation, Budget, Welfare, Subsidy, Sanction, Diplomacy, Treaty, Alliance, Conflict, War, Peace, Negotiation, Mediation, Arbitration, Resolution, Protest, Movement, Rights, Freedom, Liberty, Justice, Equality, Equity, Diversity, Inclusion, Representation, Constituency, District, Quorum, Caucus, Majority, Minority, Coalition, Opposition, Ideology, Conservatism, Liberalism, Socialism, Communism, Capitalism, Fascism, Nationalism, Populism, Globalization, Secularism, Theocracy, Pluralism, Patriotism, Sovereignty, Autonomy, Annexation, Colonization, Imperialism, Revolution, Coup, Insurgency, Terrorism, Extremism, Radicalism, Moderation, Diplomat, Ambassador, Envoy, Consul, Attaché, Intelligence, Espionage, Surveillance, Humanitarian, NGO, Civil Society, Activism, Lobbying, Advocacy, Petition, Boycott, Strike, Rally, March, Demonstration, Riot, Insurrection, Rebellion, Guerrilla, Militia, Conflict, Warfare, Arms, Military, Soldier, Army, Navy, Air Force, Marine, Commander, General, Admiral, Veteran, Casualty, Truce, Ceasefire, Armistice, Disarmament, Nonproliferation, Treaty, Summit, Conference, Convention, Protocol, Accord, Agreement, Compact, Pact, Negotiator, Mediator, Arbitrator, Facilitator, Peacemaker, Broker, Delegate, Representative, Diplomatic, Embassy, Consulate, Mission, Visa, Passport, Immunity, Extradition, Asylum, Refugee, Migrant, Border, Territory, Region, Province, County, City, Town, Rural, Urban, Metropolitan, Capital, Municipality, District, Zone, Area, Locale, Community, Society, Population, Demographics, Census, Quota, Minority, Majority, Group'**
- *Get Embeddings and perform Spectral Clustering* : Word embedding of the chosen words is obtained by using the trained Word2Vec model, and after word embedding spectral clustering is applied to find the cluster.

(5) Visualization
- *Dimensionality Reduction* : t-SNE is applied on the subset of word embeddings for dimensionality reduction, cause the clustering is visualized in 2-dimensional form.
- *Plotting a scatter plot* : A scatter plot was created to visualize the clustering results in the reduced two-dimensional space. Each point represented a word, colored according to its assigned cluster.

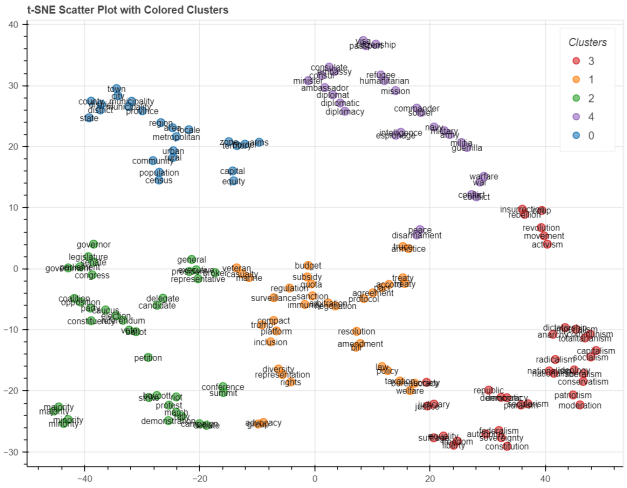

Fig. 2. Cluster of Word Embeddings from Pre-Trained Word2Vec

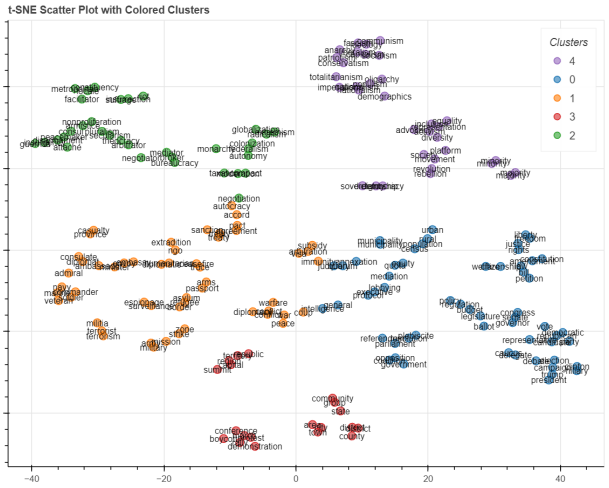

Fig. 3. Cluster of Word Embeddings from Twitter dataset Trained Word2Vec

4 VISUALIZATION

The following images are of the clusters obtained from pre-trained and the model trained by us. We used cutting-edge text analysis and visualization techniques to find subtle patterns and trends in our investigation of political bias in the Internet Research Agency's Twitter data. We set out to explore the complex terrain of political discourse found in tweets by utilizing word embeddings and clustering algorithms.

The resulting visualization shows clusters of tweets indicative of different political affiliations and sentiment polarities, elegantly capturing the multifaceted dimensions of political bias. As seen in Figure 2, the data of the pre-trained model is not politically biased as the words such as Trump, Hillary Clinton, president, and governor are not clustered together in the same group. However, on Twitter's dataset political biases are found. Words such as Trump, Hillary, President, Governor, Votes, Elections, and Legislature are clustered together in one cluster which is evidence of political biases. Figure 4 shows the cluster of word embedding with high perplexity, it shows the group and cluster of words that are closely related and hard to differentiate, thus colors are used to differentiate. Figure 5 shows the cluster in which political bias is been noticed.

In summary, this visualization is a helpful tool to uncover and understand the different aspects of political bias in Twitter data. It encourages us to think about how online discussions can influence what people think more broadly.

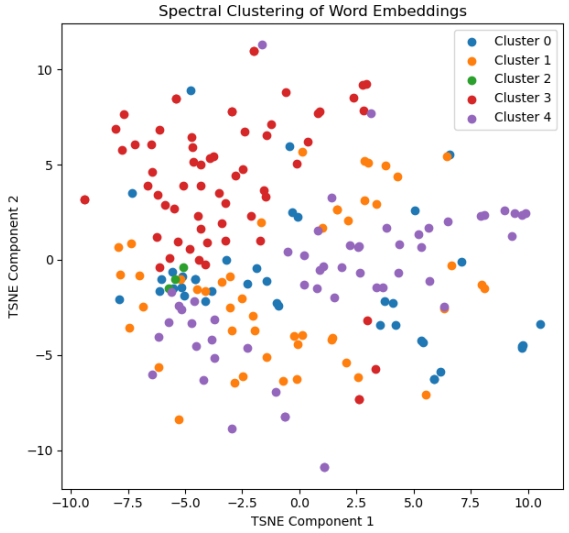

Fig. 4. Cluster of Word Embeddings with high perplexity value for Trained Twitter Russian Troll Factory Dataset

5 DISCUSSION

When we look at the pictures showing political bias on Twitter, we can see some important things. The pictures have different groups marked by colors, and these groups help us understand the complicated world of political conversations. These groups don't just show different political views; they also capture different feelings and topics. Understanding these groups helps us see that political bias is not just about putting things into two simple categories. It's much more complex and has different aspects, like how people feel and what they talk about. The pictures give us a closer look at this complexity.The effectiveness of the clustering algorithm in grouping tweets based on political bias is evident, yet certain limitations persist. The algorithm excels in capturing overt biases but may struggle with subtle nuances or sarcasm.

Comparisons between embeddings trained on our dataset and pre-trained embeddings shed light on the dataset's domain-specific nuances. While pre-trained embeddings capture broad language patterns, our dataset-specific embeddings uncover subtleties unique to political discourse. This contrast highlights the importance of

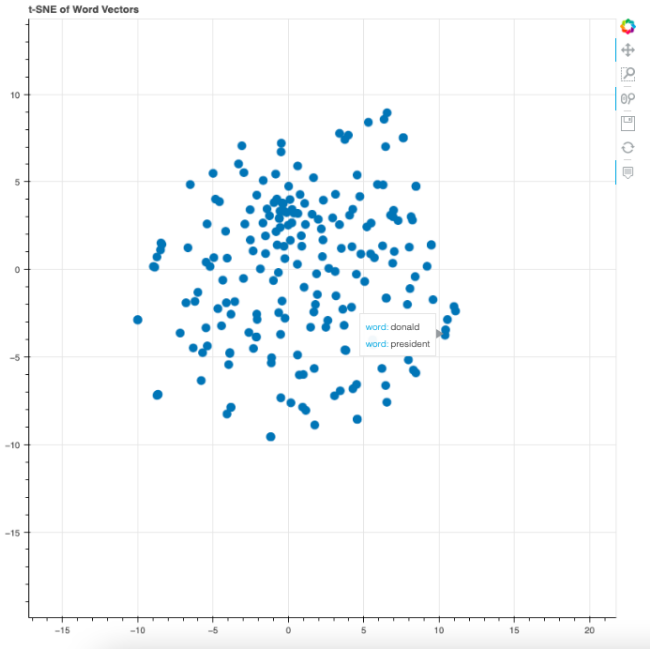

Fig. 5. Cluster of Word Embeddings showing Donald & President in the same cluster

training embeddings on contextually relevant data for a more accurate representation of political bias. This project contributes to the broader understanding of political bias in online discourse, aligning with existing literature while offering novel insights.

6 LIMITATIONS

Using vector structured data clusters, such as those generated by word embeddings offer robust ways to analyse hidden patterns including presence of bias. However, some limitations cannot be overlooked -

- *Vagueness :* The interpretation of clusters and their association with bias is subjective and depends on the analyst's perspective. There is a possibility different analysts may draw different conclusions from the same cluster.
- *Context loss :* Word embeddings are captured based on a concept. Clustering these words can lead to skewed biases or change in nature of usage.
- *Overfitting :* Clustering algorithms can lead to overfitting i.e., capturing relationships or patterns that are not relevant.
- *Contextual Ambiguity :* Words that have multiple meanings depending on the context maybe incorrectly clustered leading to improper interpretations.
- *Algorithmic Bias :* Algorithms used to induce clustering can themselves be biased.
- *Resource Intensive :* Handling vector representations can be computationally intensive and require significant memory and processing power.
- *Dependency on Corpus :* Size and quality of dataset is crucial since the vector representations are implemented on the basis of corpus used for training.

REFERENCES

[1] "Did Clinton win more votes than any white man in history?". BBC News. December 12, 2016. Retrieved September 9, 2018.
[2] Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences, 115(16), E3635-E3644.