Exploring the Armenian Political Landscape with Data Science

Yeva Tshngryan
BS in Data Science American University of Armenia Yerevan, Armenia
yeva_tshngryan.edu.aua.am

Advisor: Natali Gzraryan
American University of Armenia Yerevan, Armenia

***Abstract*—Now more than ever it is important to stay up to date with the latest political events and most importantly be able to analyze them and draw conclusions. The following paper aims to illustrate how to use data science tools when dealing with complex political situations. To make the analysis more informative and less biased, I will scrape the tweets of analysts who are not on the same political side. Some of the analyses will use the 2020 Nagorno - Karabakh war as a divider to showcase how much the engagement and sentiment changed after the war. I will also train the LSTM model on the data to give the most accurate sentiment scores. This paper aims to put the information we are getting from politicians into ordered containers, using classic data analysis techniques combined with sentiment analysis and polarity. It aims to analyze the research questions that I have defined in section 1.3.**

Key Words: Sentiment analysis - Natural Language Processing (NLP) - Politics

I. Introduction

Twitter is known for providing an opportunity for people to express their ideas using quick, short texts. This has been known to create a lot of misinformation, especially in politics. More than 50% of Twitter users discuss political issues on Twitter making it an ideal platform to analyze situations described in the study. [5] Data Science has been used in Politics to predict election results, analyze campaigns, and even use facial recognition software to monitor populations. Additionally, next to the above-mentioned use cases, there is a method in Data Science known as sentiment analysis, which aims to analyze given texts or opinions using sentiment and polarity calculations.

In recent years there has been a dramatic increase in the number of people who express their opinions online to the public, amongst these people are several political figures active on Twitter and Facebook. The latter has allowed Data Scientists to draw conclusions and make assumptions with the help of statistical, knowledge-based, and hybrid analysis approaches. The data available from people expressing their opinions on Twitter allows us to analyze the engagement and opinions of people and it can also help political journalists and analysts to understand people's opinions and the direction of their thoughts. [5]

Sentiment Analysis is a branch of natural language processing that involves analyzing the words used in a piece of text to measure its sentiment, whether positive, negative, or neutral. It's commonly used to determine how people feel about a product, a figure, or an event. This study will use sentiment analysis to analyze the political landscape of Armenia as a whole. Similar techniques have been applied to Twitter to gauge the political temperature of the platform. The focus of this paper will be surrounding the emotions of political figures, the matter of security, war, and peace, and overall events happening in the country. The study will also provide classification and clustering of tweets to provide an overall picture of what our politicians are focusing on the most.

On 2020 September 27, Azerbaijan launched a military attack on Artsakh, which resulted in digital warfare and high media activity from the Armenian leaders and respectively a higher engagement from the user's side. We will notice a definitive pattern if we observe the engagement before and after 2020. In Figures 1 and 2 we can see some examples from Nikol Pashinyan's and Edmon Marukyan's tweets engagement from users as the number of posts as well as number of likes have skyrocketed. The 2020 Artsakh war will also serve as a divider tool to observe how the social media presence of politicians changed after it.

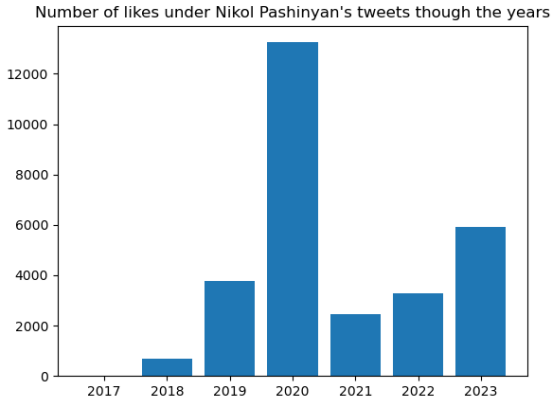

Fig. 1. Number of likes Under Nikol Pashinyan's tweets.

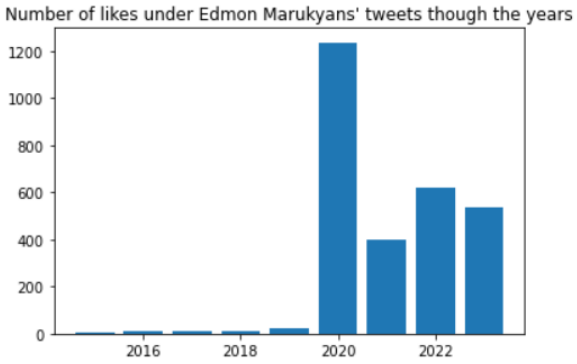

Fig. 2. Number of likes Under Edmon Marukyan's tweets.

II. Data

The data for this paper was gathered from Twitter using an extension of Pythons' snscrape library called sntwitter.

To avoid bias, the data was also collected from accounts of users who expressed a neutral opinion about every situation as well as data from different political parties with a very specific outlook on situations. Note that some political figures started being active on Twitter later than others. I have also wanted to scrape data from Facebook however due to their terms and conditions could not do it, which is why the research is mainly based on Twitter data.

The dataset contains columns, the Date Created, the Number of Likes, the Source of the tweet, and the tweet itself. Later I used the TextBlob library that calculates the polarity and later added the label based on the tweet polarity.

III. Research Questions and Hypotheses

The following study aims to find answers to the following research questions:

- What are the most frequently used words of politicians and how do they differ from time to time
- What are the main points and sentiments of different political figures
- How are the standpoints alike

Apart from research questions, there are a few hypotheses as well: The following study aims to find answers to the following research questions:

- Sentiment towards the government in Armenia is influenced by external factors, such as economic conditions and international relations, as well as domestic policies and political events.
- The main topic of discussion in the Armenian political landscape in social media is not at all centralized around Armenia, but on external topics.
- The sentiment will not be very positive overall.

IV. Methodology

Overall, the process of doing the analysis took the following shape:

1) Scraping and cleaning the data
2) Data Analysis
3) Performing Sentiment Analysis
4) Clustering with sentence transformers
5) Classification of topics with Network Analysis
6) Topic Modeling

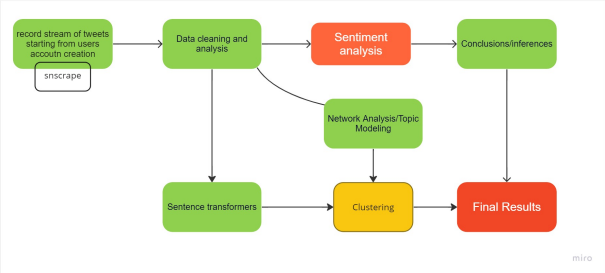

Fig. 3. Methodology of the study presented with a flowchart

Platforms such as Twitter are known for having very noisy text, with entities that are not classified as text and it can affect the sentiment scores. Many models don't take into consideration punctuation, emojis, or acronyms that's why results might not be very accurate. In order not to have the issue I have used TextBlob, which is a pre-trained model for sentiment analysis that outputs labels - negative, positive, neutral, and also sentiment scores ranging from -1 to +1, with -1 being negative, closer to 0 is neutral and closer to +1 positive. TextBlob incorporates punctuation, slang word, and emojis into the analysis.

The study will also use the LSTM model for sentiment analysis. LSTM or Long Short-term memory is a neural network commonly used in NLP tasks such as sentiment analysis. It uses various Neural networks, that in the context of sentiment analysis are trained on a large dataset and labeled as negative, positive, or neutral. The LSTM takes words and analyzes them one at a time, while simultaneously updating the internal state based on information received. Once the entire sequence has been processed, the LSTM outputs a prediction for the sentiment of the text.

The LSTM model has a separate and more detailed section.

V. Related Work

Some papers have previously utilized data science to analyze presidential elections or the overall political situation. The authors have used Machine Learning approaches, such as sentiment analysis to analyze the behavior of ruling parties or the overall situation in the country. (note: there is no such analysis done for RA, all the papers mentioned concern other countries).

Smith, J., Doe, J., and Johnson, K. (2021) have used the R libraries such as CRAN which allows them to analyze the emotions of the tweets. They have mainly collected data from the information about the 2016 US presidential elections. [1] Miguel G. Folgado and Veronica Sanz explored the democratic and republican landscapes of Spain in their 2022 article. In their work, they have considered tweets from leaders as well as political parties and also mapped the analysis and timeline according to the happening events which are what this paper

will showcase as well. [2]
Rao Hamza Ali, Gabriela Pinto, Evelyn Lawrie and Erik J. Linstead published an article in 2022 using large-scale data to do sentiment analysis on the 2020 US presidential elections. The analysis has made use of nearly 8 million tweets including deleted, suspended, and blocked tweets to be able to do the most accurate sentiment analysis possible. [3]

VI. Results and Discussion

The 2020 attack on Nagorno-Karabakh resulted in a significant contextual change in politicians' tweets. Let's take as an example the Armenian PM's tweet which has changed significantly over the last couple of years. In order to visualize the context of tweets in the most effective way, we will be utilizing the WordClouds method with Python. In order for it to be cleaner there is a defined list of stop words that were eliminated during the analysis.

Fig. 4. WordCloud of Nikol Pashinyan's tweets

Nikol Pashinyan's tweets as we can see are way more focused on Armenia and the Armenian people, with positive words such as will, people, Artsakh and naturally Azerbaijan.

Fig. 5. WordCloud of Edmon Marukyan's tweets

On the contrary, we can observe that Edmon Marukyan's tweets are heavily focused on Azerbaijan and then on Armenia and the Armenian people. We can also make an observation that he does not write about Arstakh that actively.

Ararat Mirzoyan's social media presence is mainly focused on Armenia and its' people. He writes about Azerbaijan and Artsakh in the same capacity. As expected he writes and retweets from MFA tweets, hence the MGCC abbreviation. We can also notice that he mentions OSCE group the most

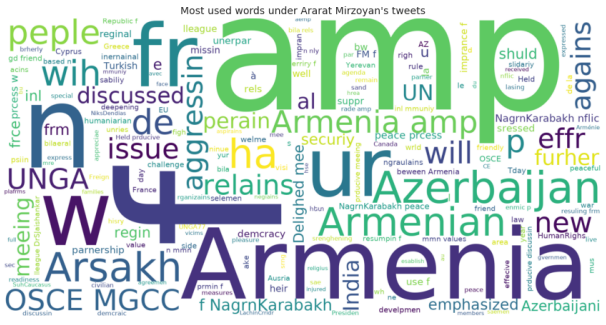

Fig. 6. WordCloud of Ararat Mirzoyan's tweets

out of all politicians.
Robert Kocharyan on the other hand has a completely different picture as he is not focusing on Azerbaijan a lot. On the contrary, he mostly includes his name in tweets, as well as talks about Armenia, Republic and mentions the words Leader and President. There are not many mentions of Artsakh as well.

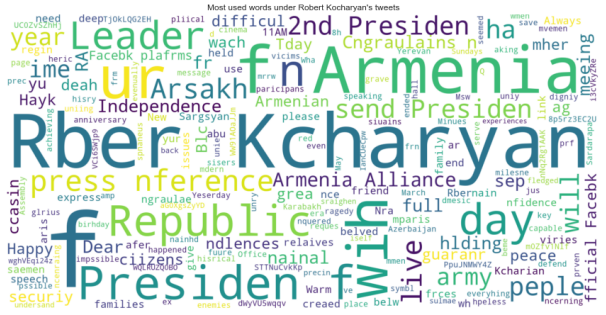

Fig. 7. WordCloud of Robert Kocharyan's tweets

Overall, the whole picture of the Armenian political landscape has the following shape of being heavily focused on Armenia and Azerbaijan, as well as Artsakh or Nagorno Kharabakh. It is interesting to notice the positive sentiment of politicians overall, with a focus on the international community and peace, as well as human rights.

VII. Polarity and Sentiment

To be able to perform sentiment analysis on the tweets, first, it is needed to determine the polarity of the tweets and according to that give them proper labels.
As mentioned earlier TextBlob has a pre-trained feature to extract sentiment from text and provide proper classification. I have combined the tweets of political leaders into one data frame to be able to extract one whole picture of what the landscape looks like. The polarity score returned by TextBlob can be used to detect the overall sentiment of a piece of text. A score of 0 indicates that the text is neutral, while a score closer to -1 indicates a negative sentiment and a score closer to +1 indicates a positive sentiment.
For the following study, I wanted to visually showcase how the sentiment of Armenian politicians changed over the years,

considering all the events happening in the Republic of Armenia, so I decided to visualize it with a bar graph of sentiments over time. Interestingly enough we can see a huge spike in positive sentiment starting from 2019 and especially 2020, if we observe figure 8. One drawback of all sentiment analysis models is that they recognize Armenian text as neutral, so it can make the analysis a little bit inaccurate.

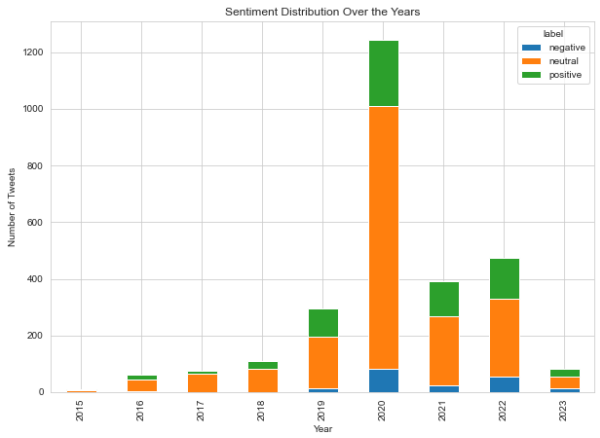

Fig. 8. Sentiment distribution of all tweets combined

If we observe Prime Minister's tweets' sentiment timeline in Figure 9, for example, we can notice a positive spike of tweets starting from 2020 and mainly 2021.

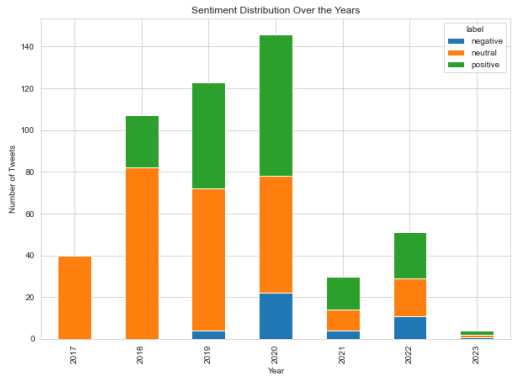

Fig. 9. Sentiment distribution of Nikol Pashinyan's tweets

I was also interested to compare the sentiment of all politicians right before and after the 2020 Nagorno-Karabakh war, so I created two variables **bold** beforeall and **bold** afterall that divide my data frame into two parts: before September 27, 2020 and after.

In figures 10 and 11, it is very interesting to observe that the overall sentiment of politicians has not been overbearingly negative even during the escalation of the situation. We can notice a negative spike in 2020 and 2022, explained by the escalation, however, overall the positive sentiment is far more prevalent in the picture.

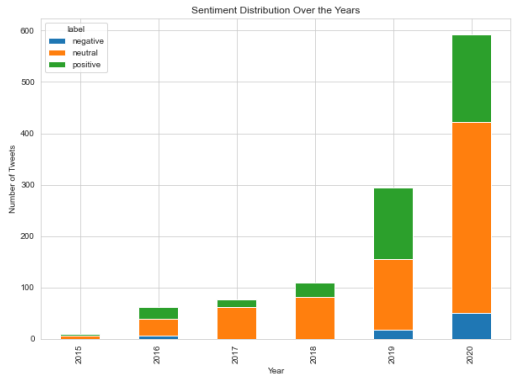

Fig. 10. Sentiment distribution of all tweets before September 27

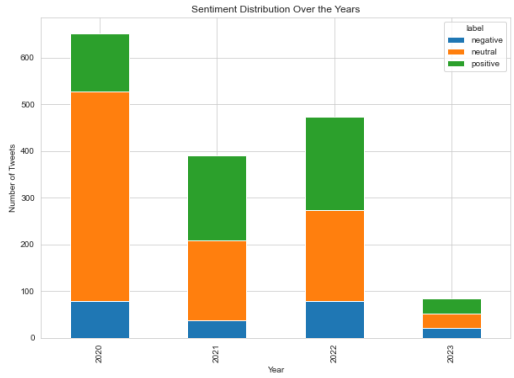

Fig. 11. Sentiment distribution of all tweets after September 27

A. *LSTM model*

To provide a more accurate sentiment analysis the study also utilizes a Long Short-term memory or LSTM, which is a type of neural network widely used for language processing tasks such as sentiment analysis.

LSTM can capture sequential data and capture long-term dependencies of it. This makes them particularly effective for modeling the relationships between words and phrases in text data, which is important for accurate sentiment analysis.

In the context of my Twitter data, LSTM can be used in various ways. I have utilized bidirectional LSTM, convolutional LSTM, and single LSTM, just to test and see which works best. As it turns out BiLSTMs are very good at capturing contextual information because it processes both forward and barckward. This can be beneficial for sentiment analysis of tweets as it can help capture the nuances and sarcasm often found in social media posts.

Single LSTMs architecture is relatively simpler and less expensive than BiLSTM and therefore the analysis is not as accurate with complex sequences.

1D LSTM is the best of both worlds, as they can extract features from input data and also capture dependencies. This makes them ideal for sentiment analysis problems.

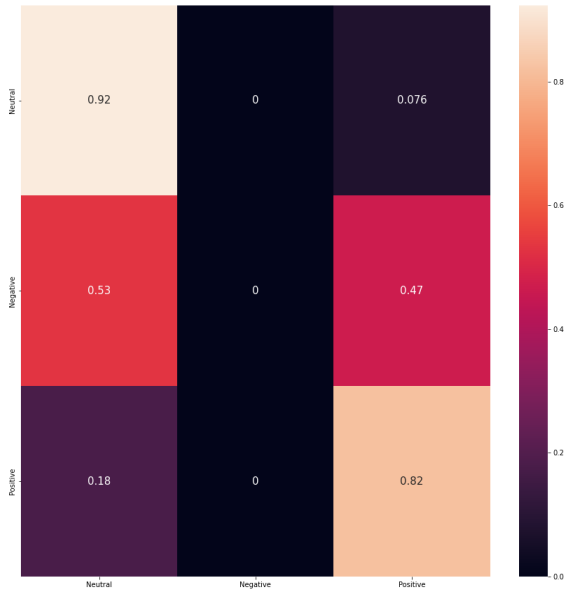

Fig. 12. Confusion matrix of LSTM model

It appears that our model is better at predicting the positive sentiment correctly as opposed to negative or neutral. This could be due to the fact that, as mentioned before, these models are not very good at detecting the sentiment of the Armenian language.

VIII. Sentence Embeddings

In order to showcase how alike Armenian politicians talk about certain problems or generally how many similar ideas they share, the study utilizes a very interesting method called sentence embedding.

In NLP, sentence embeddings are well-known technique for showing text as a vector in high dimensional space, and these embeddings can capture the meaning of the text by mapping it to a vector. The following gives us the ability to classify and cluster given data. Recently, pre-trained models such as BERT and GPT have shown impressive results in natural language understanding tasks. However, these models can be computationally expensive and require significant computational resources to train and fine-tune.

This study utilizes Tensorlow Hub, which is a platform where embeddings and pre-trained models can be shared and it provides a wide range of models for sentence embeddings that can be integrated into ML pipeline. The training is done on large datasets and optimized for specific tasks, such as sentiment analysis.

A well-known example of such a model is Universal Sentence Encoder or USE, which uses deep neural network architecture to encode text into vectors. The model is pre-trained and uses large-scale text corpus such as social media pages, news, books, etc. The study utilizes it for clustering and classification of tweets given and gives us the similarity of those.

For the specific problem of topic modeling, I wanted to take all tweets of politicians combined, because as seen in the WordClouds earlier we have quite a similar theme in politicians' tweets and as we can see from Figure 13, the hypothesis seems to be true.

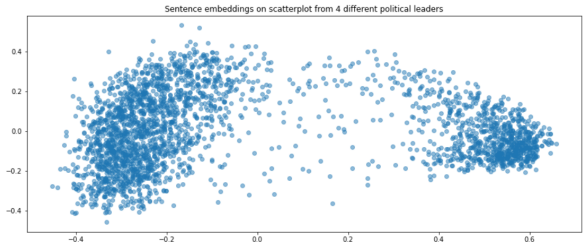

Fig. 13. A sentence embedding scatterplot of all tweets combined

The study showed that we have a pattern of our embeddings being scattered on the right and left sides of the plot. We can interpret it in a way that embeddings captured a big range of semantic meanings, with two main clusters. This means that Armenian politicians tend to write in a quite similar fashion and about the same topics. The scatterplot also revealed that there were some outliers in the data, which could tweet that does not fit into any specific cluster or are semantically different from the other tweets.

This scatterplot pattern can provide insights into the topics and themes being discussed by the political leaders on social media, as well as the relationships between these topics. The scatterplot can also be used to identify any anomalies or outliers in the data, which can be further analyzed to understand their significance.

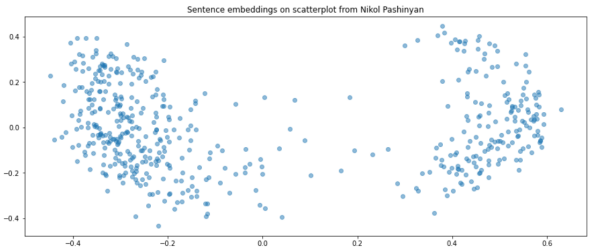

Fig. 14. Visual of Nikol Pashinyan's tweets embeddings

I also wanted to take the embeddings of politician's tweets separately, because it provides a view of their social media activity and also we can see which of the politician's really "shined through" with their topics frequency and also which of them tend to write in a more clustered manner. As we can see the Prime Minister Nikol Pashinyan tends to not spread a lot with topics and we see a similar pattern as in Fig. 13.

From the Fig. 15 we can see that Edmon Marukyan is the one who influenced the visual of Fig. 13. As we can see he tweets most frequently out of the batch and also is very concrete with topics when he does.

From Ararat Mirzoyan's and Robert Kocharyan's scatterplots we can deduce that their overall contribution to the picture is not that noticeable, as the tweets are not very concrete and are spread all over the place.

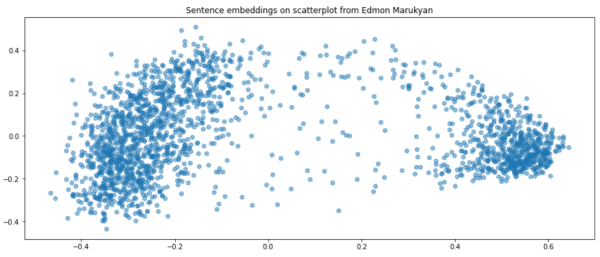

Fig. 15. Visual of Edmon Marukyan's tweets embeddings

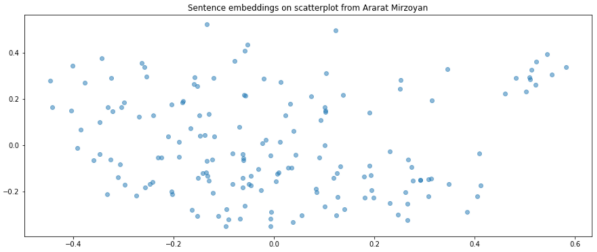

Fig. 16. Visual of Ararat Mirzoyan's tweets embeddings

However, from these plots alone we cannot get the range of all clusters and classes of topics discussed in the tweets and that is why the study utilizes Network Analysis.

A. Network Analysis

In order to see topic clusterings in more detail, the study uses Network analysis, which revealed topic classifications of politicians' tweets separately. By examining the most common words used in these clusters, the study identified the most prevalent topics discussed by each politician. Network analysis used in this context can serve as a powerful tool for researchers to gain a deeper understanding of the complex world of politics and can be used to have a more nuanced understanding of the landscape as a whole.

The insights gained from the analysis of social media data can be applied to a wide range of political data, including speeches, news articles, and public statements, making network analysis a valuable technique for researchers in the field of political analysis.

We can observe that the Network Analysis graph looks like a web, where each node represents a tweet and the edge represents how they are connected to each other. I defined a set of stop words that were removed so that they don't cause too much noise and contextual change in the analysis.

We can see very obvious clusters formed in our Network graphs, where the most prevalent topics are present. After the plotting, I extracted the most commonly used words of each Politician.

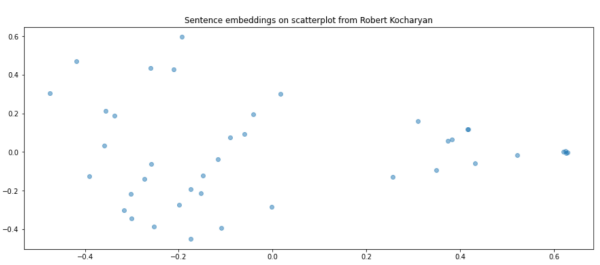

Fig. 17. Visual of Robert Kocharyan's tweets embeddings

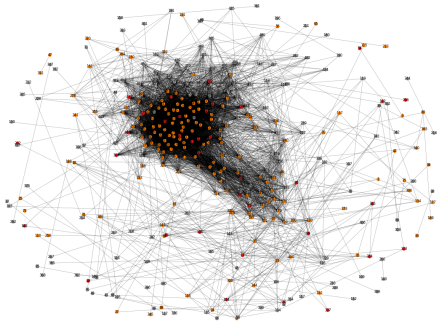

Fig. 18. Network Analysis of Nikol Pashinyan's tweets

The most prevalent topics in Nikol Pashinyan's tweets include the words: *Armenia, we, Pashinyan, Nikol, will, President.*. From this, we can conclude that Nikol Pashinyan puts a strong emphasis on the collectivity and integrity of the Armenian people, as well as readiness.

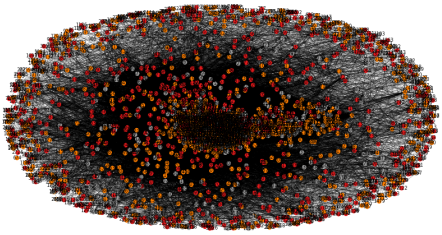

Fig. 19. Network Analysis of Edmon Marukyan's tweets

The most prevalent topics in Edmon Marukyan's tweets include the words: *We, Armenia, Today, Armenian, are, well, BrightArmenia.*. We can deduce that a lot of his focus goes into his political party, however, the notes of collectivity are still present. We can also see that the topics of his tweets do not seem to be as scattered all over the place as for example Nikol Pashinyan's tweets.

The most prevalent topics in Edmon Marukyan's tweets include the words: *Armenia, Cyprus, Security, Armenian, we, cooperation, people, congratulations, appreciate, and Turkish*. It is interesting to observe that the model picked up words such as Turkish, congratulations, and cooperation together as a connected topic throughout his Twitter activity. We can also observe very positive sentiments in tweets with words such as we, cooperation and appreciation.

The most prevalent topics in Edmon Marukyan's tweets include the words: *Robert, Kocharyan, Armenia, republic,*

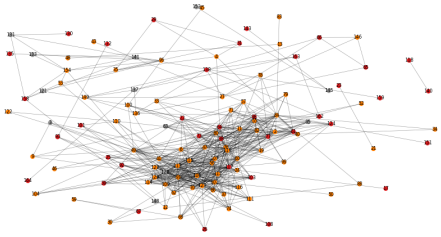

Fig. 20. Network Analysis of Ararat Mirzoyan's tweets

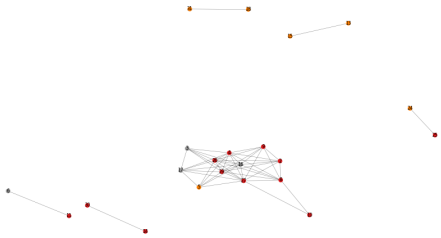

Fig. 21. Network Analysis of Robert Kocharyan's tweets

we, president, Artsakh, second, people. The model picked up almost the same theme as expected and as we observed before in the WordClouds. With Robert Kocharyan's tweets, it is expected as he started being active on Twitter very recently, so there is not too much that can be different from the earlier analysis. Because Network analysis works in a way that finds patterns and themes inside a large text corpus, I expected his results to not be different.

IX. Topic Modeling

The last technique that the study will utilize is Topic Modeling, which is a powerful data science tool allowing us to extract topics and themes given large-scale textual data. The implementation we will be using is the Latent Dirichlet Allocation or LDA algorithm, which is an extension of the Gensim library. Using LDA and the data available, the algorithm allowed us to extract 10 different underlying topics of politicians in Armenia.

The probabilistic formula behind the topic modeling process is based on the assumption that each document is a mixture of different topics, and each topic is a distribution over the words in the corpus. More formally, the topic modeling process can be expressed using the following probabilistic model: [4]

1) For each topic k, draw a distribution over words ϕ_kk Dirichlet(β), where β is a hyperparameter that controls the sparsity of the topic-word distributions.
   $\phi_k \sim$ Dirichlet(β)
2) For each document d, draw a distribution over topics θ_dd Dirichlet(α), where α is a hyperparameter that controls the sparsity of the document-topic distributions.
   $\theta_d \sim$ Dirichlet(α)
3) For each word wi,j in document di, where j is the position of the word in the document, draw a topic assignment zi,j Multinomial(θ_di) and draw a word wi,j Multinomial($\phi_k$zi,j), where Multinomial($\cdot$) denotes a multinomial distribution.
   $z_{i,j} \sim$ Multinomial(θ_i)
   $w_{i,j} \sim$ Multinomial($\phi_{z_{i,j}}$)

Applying topic modeling to our data resulted in 10 different topics, with 2 of the topics being the most commonly discussed as we can observe in the WordClouds below.

Fig. 22. Two most prevalent topics according to Topic Modeling algorithm

X. Limitations and Real-Life Use Cases

The following study has its own limitations and real-life use cases, but also some further improvements that should be considered as well.

1) **Data Sources**: The library snscrape gives a limited possibility to scrape accounts, because of its API Limitation that only allows a certain amount of accounts to be scraped at a time. On the contrary, there is another possibility to scrape data with the Tweepy library, which unfortunately needs permission from the Twitter Developer account, where I submitted an application to acquire Twitter API Consumer Key, but never heard an answer.
2) **Language Limitations**: libraries such as LSTM and TextBlob do not have a built-in feature to be able to label the Armenian language, so it labeled all unknown words and text as "neutral" which makes the insights not entirely thorough.
3) **Model Limitations**: the LSTM, networkX, and TextBlob models might not be able to catch irony and sarcasm in the texts and obviously can't analyze the

meaning behind Armenian texts. There might be an issue with data clustering because Armenian texts might just become outliers in the visuals.

As for real-life use cases, there are a few possibilities:

1) **Political strategies**: With the analysis political candidates can develop better campaign strategies and be able to analyze public opinions better.
2) **Media Analysis**: Political analysts or journalists can identify patterns of behavior of politicians using the analysis because they cannot deal with so much data manually.

For further improvements, I have decided to start working on building a sentiment analysis model for the Armenian language as it can be a great tool for reducing bias and discrepancies in analysis as such and can help to analyze the sentiment and polarity of the Armenian side of social media. Unfortunately, currently, there is not much-structured data available for an analysis like this in the Armenian language so an improvement like this might take a longer time to gather social media data only in Armenian.

XI. Conclusion

In conclusion, we can deduce that Armenian politicians' tweets that the study analyzes have quite similar styles and semantics when it comes to countries' internal and external issues. It was notable seeing how the positive sentiment grew after the 2020 war and is persistent even now.
However, to get the full picture of the landscape it is also essential to be able to analyze tweets in other languages as well, in this case in Armenian, as we saw in the analysis Armenian tweets get neutral sentiment scores and it can be hard to analyze the tweets' whole picture in that case.
The study used 3 different methods to analyze the most relevant topics in politician's tweets and there were some interesting and different insights that emerged. Word Clouds, Topic modeling, and Network analysis are 3 different methods to provide perspective on data, and they analyze a corpus of text in different ways.

- **Word Clouds**: Word Clouds is a simple method to visualize the most frequently occurring words. The most frequently occurring words are presented with larger fonts, and less and less occurring ones in smaller ones. It is a quick and easy way to get an overall idea of what we are dealing with.
- **Network Analysis**: Network analysis is an interesting method as it identifies relationships between words in a large corpus of text and we can visualize it via a web network. The results we got from Network Analysis are different than the ones from Word Clouds, as it identifies the frequent occurrence of words based on the context in which it is present.
- **Topic Modeling**: Topic modeling is more of a statistical method to identify latent topics and themes in the text. It identifies patterns of occurring words and groups them together in separate topics. Each topic represents a set of words that happen to occur together more frequently. It is a useful method if we want to get a more nuanced idea of our text data and get underlying themes within it.

We have disproved the hypothesis that the main focus of Armenian politicians concerns the external situation and is not very "Armenia - focused". As we saw in the Network Analysis the main theme of all politician's tweets was centered around Armenia and Armenian people. It is interesting to see that Topic modeling picked up a theme of Azerbaijan being mentioned in context of Russian tweets.

XII. References

[1] Smith, J., Doe, J., and Johnson, K. (2021). Analyzing Political Sentiment Using Twitter Data. Journal of Social Media Analysis, 5(2), 45-62.

[2] Folgado, M. G., and Sanz, V. (2019). Exploring the political pulse of a country using data science tools. Social Science Computer Review, 37(5), 616-632.

[3] Ali, R. H., Pinto, G., Lawrie, E., and Linstead, E. J. (2021). A Large-Scale Sentiment Analysis of Tweets Pertaining to the 2020 US Presidential Election. In 2021 IEEE International Conference on Big Data (Big Data) (pp. 2562-2570). IEEE.

[4] Akhmedov, F., Abdusalomov, A., Makhmudov, F., and Cho, Y. I. (2021). LDA-Based Topic Modeling Sentiment Analysis Using Topic/Document/Sentence (TDS) Model. Entropy, 23(6), 685. doi: 10.3390/e23060685

[5] McMinn AJ, Moshfeghi Y, Jose JM. Building a large-scale corpus for evaluating event detection on twitter. In: Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, 2013; pp. 409–418.