Abdul Saboor Hamedi

**Introduction**

Data Afghanistan is controlled by the Taliban, and this group seized the power on August 15, 2021, the Taliban well known as an extremist Islamic group, they have left over a very harsh experience back in 1996. Since then, there have tons of research about this group, some used secondary data while some could make primary data through interviews. Today we have seen that the Taliban did not change and still carries that extremist ideology, especially regarding women, lately the Taliban banned to go schools, and this action caused the international community and other communities to denounce them. Today no girl can get an education, because the schools are closed for girls, and half of the Afghan society has been disabled through the Taliban ideology, thus I have launched this study to find out their ideologies toward women, especially in areas of education and politics.

This study will modern approach using Natural Language Processing (NLP) is an area within artificial intelligence and is referred to as the process to read, analyze, and understand a large amount of text data and is considered to be the future for understanding the unstructured text. Twitter data were obtained through Twitter’s open API. Data from three parks in Stockholm were collected between the periods 2015-2019. Three analyses were then performed, temporal, sentiment, and topic modeling analysis. The results from the above analyses show that it is possible to understand what attitudes and activities are associated with visiting parks using NLP on social media data.

It is clear that sentiment analysis is a difficult task for computers to solve and it is still in an early stage of development. The results from the sentiment analysis indicate some uncertainties. To achieve more reliable results, the analysis would consist of much more data, more thorough cleaning methods, and be based on English tweets.

One significant conclusion given the results is that the Taliban’s ideology would be identified based on their tweets, this means we can find their ideologies individually, whether their ideologies correlated to the other Taliban or not. This research would use three different tweets, 1) belongs to the activists and journalists, what they say about the Taliban, I am doing this because they get updated with daily information, what the Taliban doing, and what policy they have implemented. 2) People, people play an important role in every society because every policy would affect people’s daily life, thus this study would collect data in general at this stage. 3) The Taliban themselves, this study also would collect data through the Taliban’s Twitter, and this would be the most important data, where they upload every policy and regulation.

**Research Question and variables**

Question: What is the Taliban's ideology toward women's politics and education?

**Variables:**

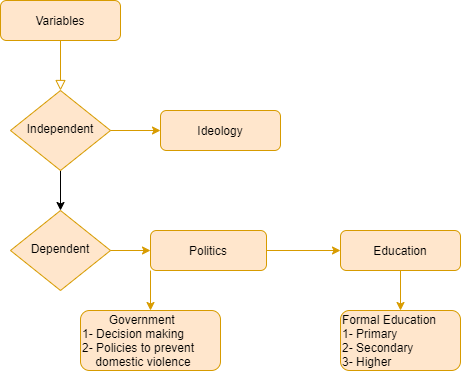


Figure 1.0: ID & DP variables

Here are the variables, the independent variable is ideology, which explains followed by the dependent, this study has two DP, which are, politics and education, also the DP has variables, the politics would be decision-making and policies to prevent domestic violence, and the education would the formal education such primary, secondary and higher educations.

**Approach**

In this report, all data gathering, and analysis will be done using Python. The reason for this is that Python contains many suitable libraries for data gathering as well as Natural language processing. The following approach will be taken to answer the research questions: First, the Twitter data will be obtained through Twitter’s open API which enables third parties to access Twitter data. After the Twitter data has been collected and preprocessed the temporal, sentiment, and topic modeling analysis will be performed. Two approaches are available when performing sentiment analysis, rule-based approach (lexicon-based) and automatic approach.

**Natural Language Processing**

Natural Language Processing often referred to as NLP is an area within artificial intelligence and can be explained as the process to read, analyze, and understand a large amount of text data. The objective of NLP is to understand human language through reading text and drawing valuable insights from it. Today NLP is considered a tough problem to solve for computers due to the nature of the human language. Computers must both understand the meaning of words and how they are used in a sentence to express the intended message in order to completely understand human language. Today billions of texts are being generated daily on various social media platforms such as Facebook, YouTube, Instagram, and Twitter. All this text data is generated every day in an unstructured form which makes traditional approaches to interpreted text data time- and resource-consuming and no longer an appropriate choice. This is where the importance of NLP appears with the ability to automatically deal with large amounts of unstructured text.

With the increasing amount of data being generated every day the use of NLP will be only become more and more important and it considered to be the future of understanding text data. Already today NLP is used in a variety of applications from which we interact daily from translate applications such as Google Translate to speech recognition applications such as Amazon Alexa or Google Home.

**Sentiment analysis**

Sentiment analysis is an area within Natural Language Processing that refers to the process of automatically obtaining emotions in texts and classifying their polarity as either positive, negative, or neutral. Standard sentiment analysis works as such that it reads each word separately in a text, classify it according to its semantic orientation and analyze how each word’s polarity affects the other and gives a collected sentiment score of the whole text.

Sentiment analysis is considered a difficult problem for computers to solve since texts can consist of several underlying sentiments, so the computer needs to understand the intended message in order to successfully classify its semantic orientation. Two approaches are available when performing a sentiment analysis, rule-based or automatic approach. The rule-based method uses a set of linguistic rules to identify the polarity in text data. The rule-based method is often implemented through a lexicon consisting of a list of words that have been labeled according to their semantic orientation as either, positive, negative, or neutral.

The automatic approach, on the contrary, does not rely on rules instead it uses machine learning techniques. Sentiment analysis is a task that can be modeled as a classification problem. The model takes a given text as input and returns a classified sentiment of the whole text as either positive, negative, or neutral as output. This works by feeding the model with classified text data and thus training the model to associate a given input to a corresponding sentiment.

**Challenges of Sentiment analysis**

Today sentiment analysis is not able to be as precise with its classification as humans. Understanding of how words affect each other in a text to convey the intended sentiment is one of the challenges for sentiment analysis. The use of informal language such as slang, emoticons, and conjunctions is another challenge to interpret for sentiment analysis. About 97 percent of all written text on social media contains informal words. With the increasing amount of data being generated every day sentiment analysis will only become more and more important. It enables companies, institutions, and researchers to automatically analyze a large amount of unstructured text data, and continued development in this area is of great importance.

**VADER**

VADER is an abbreviation for Valence Aware Dictionary and Sentiment Reasoner and is an open-source lexicon rule-based sentiment analysis model specifically adapted to detect sentiments in social media texts. The VADER lexicon is a collection of several sentiments lexicons consisting of more than eleven thousand words and characters where each item has been given a score after its semantic orientation as either positive, negative, or neutral. Since VADER is specifically created for dealing with social media texts it does not only classify words it also classifies, slang, acronyms, punctuation, and emoticons (short for emotion icon) according to their semantic orientation [12]. When determining the sentiment of text, in addition to only classifying it as either positive, negative, or neutral VADER also tells the magnitude of polarity. The output from VADER’S sentiment scoring consists of four metrics, positive, negative, neutral, and compound. Each of the metrics positive, negative, and neutral represents the proportion of the text that falls into one of these categories. The compound score is calculated by summing all the lexicon ratings for each word in the text and standardizing its value between -1(most negative) to +1(most positive). The compound score is the most useful metric if one is to use only one metric to determine the sentiment of a given text and is the metric used in this report when classifying texts.

**Data collection**

The first step consists of obtaining data from Twitter, this section describes how the data was collected from Twitter, followed by a description of the data obtained for each politics and education.

**Twitter data collection**

Through integration with Twitters open API data from Twitter could be obtained. All data were collected between the period 2021-2022. The Twitter data was obtained for each dependent variable with the use of a conditional statement including Keyword, and hashtag (#).

**Data Preprocessing**

The data preprocessing phase aims to remove irrelevant data and is a critical stage to obtain a high accuracy from the analysis. After all, tweets were collected for each park, the metadata objects “created\_at” and “text “was extracted. The “created\_at” and “text” attribute represents the UTC time when the tweet was created and the actual text tweet respectively. Below follows a description of the method used to clean the data for each analysis, temporal, sentiment, and topic modeling.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Timestamp | Year | Month | Day | Time |
| 2021-10-13 15:54:07+00:00 | 2021 | 10 | 13 | 15:54:07 |
| 2021-12-18 00:20:41+00:00 | 2021 | 12 | 18 | 0:20:41 |
| 2022-10-11 13:53:22+00:00 | 2022 | 10 | 11 | 13:53:22 |
| 2021-09-26 07:04:58+00:00 | 2021 | 9 | 26 | 7:04:58 |
| 2022-08-21 19:22:35+00:00 | 2022 | 8 | 21 | 19:22:35 |

Table 1.0: Data frame showing the top five tweets

**Sentiment analysis**

The data preprocessing for the sentiment analysis consisted of first extracting the metadata objects “created\_at” and “text “which were then loaded into a Python data frame.

|  |  |
| --- | --- |
|  | Post |
| 1 | 454 days since the Taliban banned teenage girls from school. Afghan girls continue to be locked out of their classrooms and denied their basic human rights #LetAfghanGirlsLearn |
| 2 | 453 days since the Taliban banned teenage girls from school. Afghanistan remains the only country in the planet banning children from getting an education #LetAfghanGirlsLearn |
| 3 | RT @BBCYaldaHakim: "I'd like to spend 100% of my energy to stopping executions in Iran. Now I have to deal with this mess which takes a lot… |
| 4 | 451 days since the Taliban banned teenage girls from school. Afghan girls have been forced to contemplate a life without formal education and the world remains mostly muted #LetAfghanGirlsLearn |
| 5 | "I'd like to spend 100% of my energy to stopping executions in Iran. Now I have to deal with this mess which takes a lot of energy of the European Parliament. This is a self made problem. The real problems, I don't want them to fall down priority agenda" German MEP @HNeumannMEP https://t.co/UrkgHZpZzJ |

Table 1.1: Sentiment data frame after loaded into python, showing ten selected tweets.

**Clean Data**

|  |  |
| --- | --- |
|  | Post |
| 1 | 454 days since the Taliban banned teenage girls from school Afghan girls continue to be locked out of their classrooms and denied their basic human rights LetAfghanGirlsLearn |
| 2 | 453 days since the Taliban banned teenage girls from school Afghanistan remains the only country on the planet banning children from getting an education LetAfghanGirlsLearn |
| 3 | I’d like to spend my energy on stopping executions in Iran Now I have to deal with this mess which takes a lo |
| 4 | I would like to spend my energy on stopping executions in Iran Now I have to deal with this mess which takes a lot |
| 5 | 451 days since the Taliban banned teenage girls from school Afghan girls have been forced to contemplate a life without formal education and the world remains mostly muted LetAfghanGirlsLearn |

Table 1.2: Sentiment data frame after loaded into python, showing ten selected tweets, clean data.

**Most Common Words**

|  |  |
| --- | --- |
| Words | Count |
| Girls | 1283 |
| Taliban | 1144 |
| Afghanistan | 984 |
| School | 618 |
| Afghan | 499 |
| Education | 444 |
| Amp | 423 |
| Letafghangirlslearn | 417 |
| Teenage | 403 |
| Women | 368 |

Table: 1.4 the ten most frequent words count

Here are the ten most frequent words which are used on Twitter by the user. The word girl used in 1283, followed by the Taliban which is 1144, and so on.

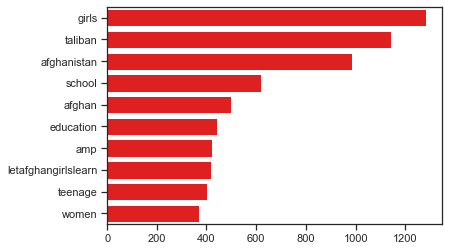


Figure 1.1, Most frequent words

And, here is graph which show the most frequent words which is been used, in the dataset.

**Tokenization**

|  |
| --- |
| Tokenize |
| [days, since, the, taliban, banned, teenage, girls, from, school, afghan, girls, continue, to, be, locked, out, of, their, classrooms, and, denied, their, basic, human, rights, letafghangirlslearn] |
| [days, since, the, taliban, banned, teenage, girls, from, school, afghanistan, remains, the, only, country, in, the, planet, banning, children, from, getting, an, education, letafghangirlslearn] |
| [id, like, to, spend, of, my, energy, to, stopping, executions, in, iran, now, i, have, to, deal, with, this, mess, which, takes, a, lot] |
| [days, since, the, taliban, banned, teenage, girls, from, school, afghan, girls, have, been, forced, to, contemplate, a, life, without, formal, education, and, the, world, remains, mostly, muted, letafghangirlslearn] |
| [id, like, to, spend, of, my, energy, to, stopping, executions, in, iran, now, i, have, to, deal, with, this, mess, which, takes, a, lot, of, energy, of, the, european, parliament, this, is, a, self, made, problem, the, real, problems, i, dont, want, them, to, fall, down, priority, agenda, german, mep] |

Table 1.5: Tokenization

The first step of the NLP process is gathering the data (a sentence) and breaking it into understandable parts (words). Here’s an example of a string of data: “454 days since the Taliban banned teenage girls from school Afghan girls continue to be locked out of their classrooms and denied their basic human rights”

**Stop Words**

|  |
| --- |
| Remove stopwords |
| [days, since, taliban, banned, teenage, girls, school, afghan, girls, continue, locked, classrooms, denied, basic, human, rights, letafghangirlslearn] |
| [days, since, taliban, banned, teenage, girls, school, afghanistan, remains, country, planet, banning, children, getting, education, letafghangirlslearn] |
| [id, like, spend, energy, stopping, executions, iran, deal, mess, takes, lot] |
| [days, since, taliban, banned, teenage, girls, school, afghan, girls, forced, contemplate, life, without, formal, education, world, remains, mostly, muted, letafghangirlslearn] |
| [id, like, spend, energy, stopping, executions, iran, deal, mess, takes, lot, energy, european, parliament, self, made, problem, real, problems, dont, want, fall, priority, agenda, german, mep] |

Table 1.6: Stop Words

The words which are generally filtered out before processing a natural language are called stop words. These are actually the most common words in any language (like articles, prepositions, pronouns, conjunctions, etc) and do not add much information to the text. Examples of a few stop words in English are “the”, “a”, “an”, “so”, and “what”. Have a look at table 1.5, these two tables are now different.

**Sentiment Analysis**

Here is the top five sentiment analysis result, this result for the top five tweets, which is in the table above.

|  |  |  |  |
| --- | --- | --- | --- |
| Positive | Negative | Neutral | Compound |
| 0 | 0.282 | 0.718 | -0.7096 |
| 0 | 0.167 | 0.833 | -0.4588 |
| 0.313 | 0.279 | 0.408 | 0.128 |
| 0 | 0.25 | 0.75 | -0.7184 |
| 0.195 | 0.311 | 0.494 | -0.4628 |

Table 1.7: Sentiment top five tweets

Each tweet was then classified into either positive, negative, or neutral based on the sentiment score given by VADER. The classification into the above categories was based using the typical threshold for classifying the polarity of texts as follows

1. Positive sentiment: Compound score >0.05
2. Negative sentiment: Compound score < -0.05
3. Neutral sentiment: Compound score > -0.05 and <0.05

**Total Of Sentiment**

|  |  |  |
| --- | --- | --- |
| Negative | Positive | Neutral |
| 1578 | 1040 | 610 |

Table 1.8: Total Of Sentiment

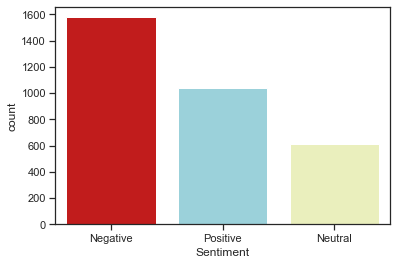


Figure 1.2 Sentiment Analysis Graph

**Word Cloud**

Word clouds (also known as text clouds or tag clouds) work in a simple way: the more a specific word appears in a source of textual data (such as a speech, blog post, or database), the bigger and bolder it appears in the word cloud. A word cloud is a collection, or cluster, of words depicted in different sizes. The bigger and bolder the word appears, the more often it’s mentioned within a given text and the more important it is.



Figure 1.3 Word Cloud