CHALMERS
UNIVERSITY OF TECHNOLOGY

Methods for detecting echo chambers in social media networks

Using Natural Language Processing to explore the differences in language use within the Swedish NATO debate

Master's thesis in Complex and Adaptive Systems

BRIAN BONAFILIA

DEPARTMENT OF PHYSICS / DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2023
www.chalmers.se

Master's thesis 2023

Methods for detecting echo chambers in social media networks

Using Natural Language Processing to explore the differences in language use within the Swedish NATO debate

BRIAN BONAFILIA

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Physics / Department of Computer Science and Engineering
Chalmers University of Technology
Gothenburg, Sweden 2023

Methods for detecting echo chambers in social media networks
Using Natural Language Processing to explore the differences in language use within the Swedish NATO debate
BRIAN BONAFILIA

Supervisor: Sebastianus Bruinsma, Department of Computer Science and Engineering
Examiner: Moa Johansson, Department of Computer Science and Engineering

Master's Thesis 2023
Department of Physics / Department of Computer Science and Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: The flag of the Kingdom of Sweden alongside the flag of the North Atlantic Treaty Organization

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2023

Methods for detecting echo chambers in social media networks
Using Natural Language Processing to explore the differences in language use within the Swedish NATO debate
BRIAN BONAFILIA
Department of Physics / Department of Computer Science and Engineering
Chalmers University of Technology

Abstract

This thesis presents an approach to using Natural Language Processing to detect echo chambers in social media networks and to find identifying terms for those echo chambers. A dataset consisting of posts and user information from the micro-blogging service Twitter related to Sweden's application to join the North Atlantic Treaty Organization was collected for the year leading up to the Swedish national election of 2022. Tight-knit communities of users on the platform were extracted using the Infomap and Leiden Algorithms based on user connections and interactions. From each community found using these methods, the corpus composed of the text postings of the users in that community was used to train a Word2Vec model to recover vector word embeddings for key words related to the subject of the discussion. Semantic change was quantified by assessing the differences in cosine similarity between pairs of words over time and between communities. Changes in the use of terms related to the subject over time were observed, but patterns representing possible echo chambers arose only with the aid of manual annotation of user positions on the issue. Conclusions could not be drawn about how successful the method is from the results alone, as evidence suggests that the issue was insufficiently polarized to generate strong echo chambers.

Keywords: word2vec, word embedding, echo chambers, community detection, polarization

Acknowledgements

I would like to acknowledge the help and guidance of my supervisor, Sebastianus Bruinsma, throughout the duration of this project. I would also like to thank Denitsa Saynova for sharing her knowledge of the Word2Vec model and for helping me find resources for the project.
Additional thanks are owed to my examiner, Moa Johansson, for her time and for her assistance in setting up the project within her department.

Brian Bonafilia, Gothenburg, Sweden 08/01/23

Contents

1

Introduction

This chapter covers the motivation for this research, the specific research questions, and some historical context for the topic that was chosen. This historical context is a general explanation of some key events leading up to the present and is intended only to give the reader enough information about the subject to understand the discussion and conclusions of this paper.

1.1 Motivation

Social media has had a profound effect on the political communication landscape in the last two decades. Dissemination of information and political campaigning via social media platforms, such as the micro-blogging service Twitter, have steadily been supplanting traditional media outlets. In Sweden today, more political communication is presented over the internet than through print media or television[1].

Because of the potential effect on democratic systems, much research has been conducted on social media communication. One major area of interest is a phenomenon known as ”echo chambers”. An echo chamber is a collection of users who communicate primarily with each other and have little outside influence on their values, opinions, or messages, resulting in a positive feed-back loop where the lack of rebuttal or disagreement leads to increasingly polarized positions. The polarization of these online communities translates to observable changes in voting patterns. It has been observed that the opinions presented by Twitter users has influenced other users' choices at the polls, with a study of the 2016 and 2020 US Presidential Elections showing that Twitter usage was correlated with a drop in support for Republican candidates due to a left-leaning tendency of Twitter postings about the election and higher popularity of the Democratic opponents on the platform[2].

This concept of echo chambers is well-known, but what is less understood is what is being echoed in them. Studies of echo chambers have analyzed the group opinion, inferring a general leaning and quantifying the echo chamber in terms of homophilly and bias[3]. But currently, there is little research into methods for identifying exactly what terms and expressions related to a polarized subject arise within and propagate through echo chambers and how the structures of the network clusters they propagate through effect the meaning and spread of the terms.

This work seeks to explore natural language processing (NLP) methods, primarily dealing with vectorized word embeddings, that can be used to identify which words and phrases have been adopted by or markedly changed meanings within an echo chamber. Terms which have changed meaning can include potential shibboleths and “dog whistles”—words consciously being used to express a hidden meaning understood only by an “in group”—as well as words or phrases which have unconsciously taken on a new meaning reflecting a change in a group's opinion. Improving the identification of these terms would have many potential uses. For example, identifying specific terms and phrases which propagate widely may help to contain the spread of online misinformation. Additionally, being able to detect terms with new use patterns reflecting deliberate changes in meaning may help to better moderate online content as extremist social media users have been bypassing

regulations against hate speech by creating their own language based on innocuous terminology[4]. A better understanding of terminology that has seen a divergence in meaning between different social network clusters can help to combat polarization by identifying where different groups are effectively speaking a different language. Finally, detection of rapid or large semantic shifts in terms can help to pin-point deliberate manipulation attempts. Prior research has indicated that the recipient being aware of an attempt to spread online disinformation make the disinformation less likely to have an effect[1].

1.2 Research Question

This research project seeks to expand on the previous research by examining two hypotheses.

1) Community groupings based on the topology of a social media network represent groups of users who will have greater similarity of opinion amongst themselves than they will with members of other user groups.

2) Language use within the community will change as a result of the differences of opinion, showing a distinction in use patterns between users who are members of one group and users who are members of another.

If both these are true, then the combined hypothesis should stand as well:

3) Community groupings based on the topology of a social network show a distinction in patterns of language use.

To explore these hypotheses, a topic that is expected to have divergent opinions and a readily-gathered dataset is considered.

1.3 Context

For this work, only a single topic is considered. This makes it possible to look at the discussion captured within the dataset from the baseline assumption that all postings are being made within the same context.

The context for this work is the Swedish discussion on Twitter surrounding Sweden's potential entry into the North Atlantic Treaty Organization. This topic was chosen because it is a public debate for which there are two major, opposed positions (one in favor of Sweden joining NATO, the other against), allowing differences of opinions to appear among the users in the dataset. It is also highly relevant to Swedes and less so for people outside of Sweden, giving narrow geographic and linguistic contexts, eliminating the concern that different semantic meaning for words would be driven by geographic or cultural differences rather than differences of opinion.

Because the discussion in the tweets focuses on several of the major concerns with Sweden joining or not joining NATO, a brief summary of the context within which Sweden is set to join NATO follows.

1.3.1 The North Atlantic Treaty Organization

The North Atlantic Treaty Organization (NATO) is an international military alliance and collective defense pact with 30 member states in Europe and North America with the stated purpose "to secure peace in Europe, to promote cooperation among its members, and to guard their freedom"[5]. Member states of NATO have standardized training and procurement procedures, conduct military exercises together, and share a collective defense obligation. Members of NATO are allied politically and militarily on the global stage, and together account for 57% of global military expenditures[6].

NATO was created with the signing of the North Atlantic Treaty on April 4, 1949. Even though the first formal NATO military operation was not until 1992, after the dissolution of the Soviet Union, NATO's focus since its inception has been to act as a safeguard against possible encroachment into Europe by the Soviets. At its creation, NATO had only twelve members: the United Kingdom, France, Luxembourg, the Netherlands, Belgium, the United States, Portugal, Italy, Canada, Iceland, Norway, and Denmark. Since 1949, 18 other European states have joined the alliance, 10 of them former members of the Soviet-led Warsaw Pact. Figure 1.1 shows the current members of NATO.

The Russian Federation, successor state to the Soviet Union, has long seen NATO as a threat and has vocally opposed further eastward expansion. Russian President Boris Yeltsin said in a 1997 meeting with US President Bill Clinton, "We believe that the eastward expansion of NATO is a mistake and a serious one at that." A US State Department memo from 1990 acknowledged the Soviet concern: "[We] do not, in any case, wish to organize an anti-Soviet coalition whose frontier is the Soviet border. Such a coalition would be perceived very negatively by the Soviets." [N1]

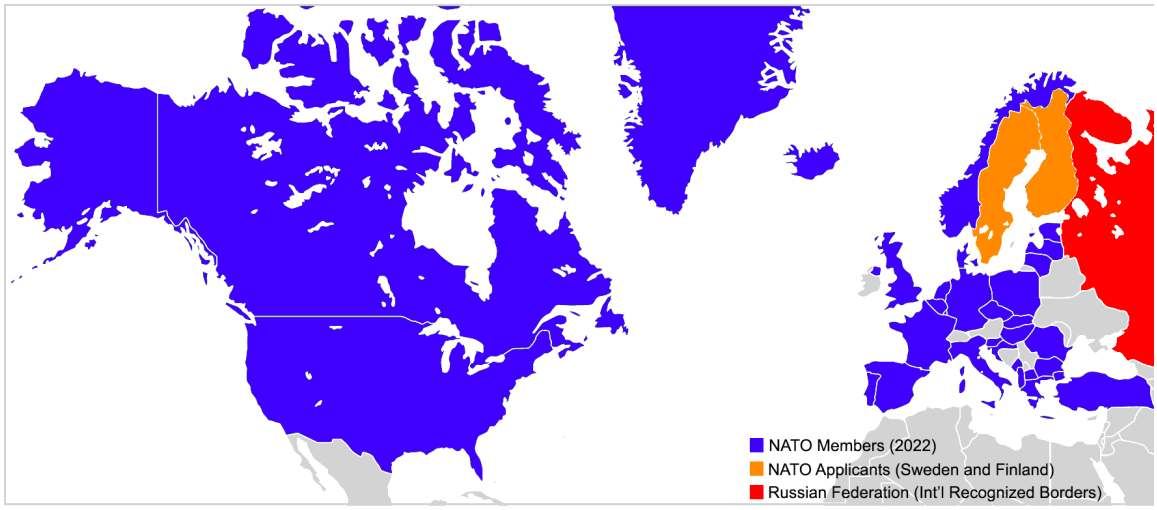

Figure 1.1: Member states of NATO as of March 2023.

1.3.2 Sweden's Relationship with NATO

Among those European nations that did not join NATO are historically neutral Sweden and it's neighbor, Finland. When NATO was formed, Sweden did not join the alliance, instead sticking to a policy of neutrality that had arisen after the Napoleonic Wars in the 19th Century and which many Swedes believed had kept Sweden out of war since. A Swedish move towards NATO was perceived as a risk to neighboring Finland, as it was expected that the Soviet Union, which exerted a great deal of control over Finland, would invade should the balance of power in the North change[N2].

Neutrality offered Sweden the ability to engage diplomatically with both sides of the Cold War, even if it meant bucking the Western trend. Sweden used its position as a neutral nation to offer rare (among the West) criticisms of US foreign policy, especially regarding America's wars in Southeast Asia in the 1960s and 1970s which lead to a chilling of Sweden-US relations[7]. Sweden was the first Western country to establish an embassy in North Korea and was the only Western country with an active diplomatic mission to North Korea until 2001[N3]. This diplomatic presence allowed Sweden to enter negotiations with North Korea, for example to achieve the release of foreign citizens held in North Korean prisons[8]. Formal membership in NATO was seen as an obstacle to these kinds of diplomatic engagements, as neutrality lent credibility to Sweden's humanitarian position.

Sweden was always West-leaning, however. Clandestine military support for Sweden in the event of a Soviet invasion was being offered by NATO[7], and after the fall of the Soviet Union Sweden joined the European Union after a public referendum.

Since the fall of the Soviet Union, NATO membership has appeared to be less relevant. Since 2015, the Center-Right parties in Sweden (Moderates, Christian Democrats, Liberals, and Center Party) had been in favor of joining NATO, while the Social Democrats, Left Party, Green Party, and Sweden Democrats have remained opposed. Generally, low public support for joining NATO has kept the issue from being a major one: in 2015, even after Russia's annexation of Crimea and increased Russian activity in the Baltic Sea, Swedes still opposed joining NATO 47%-to-33%[N4].

This low public support combined with the fear of antagonizing Russia has kept Sweden and Finland from joining the alliance. Finland shares a 1340-km border with Russia, and Swedish territory includes several strategic locations for controlling the Baltic Sea and thus Russia's access to the Atlantic Ocean. Because of this, NATO membership for Sweden and Finland has always been strongly opposed by Russia. In 2015, Russian Ambassador Viktor Tatarintsev reiterated to Swedish Foreign Minister Margot Wallström that "Swedish membership in NATO would have politico-military and foreign policy consequences, and would require retaliatory measures from Russia." [N5]

1.3.3 Events Leading to the Swedish NATO Application

In February 2014, Ukrainian President Viktor Yanukovych was removed from power following a vote in the Ukrainian Parliament after months of unrest in the country.[N6] Yanukovych had refused to sign the European Union–Ukraine Association Agreement in November 2013, which would have led to greater cooperation between the EU and Ukraine, despite majority support for measure in the Ukrainian Parliament.[N7] Yanukovych is on record as calling the deal a "humiliation" [N8] and had insisted that the EU increase its offer. His party also prevented the passing of six points that the EU had set as a requirement of signing the argument, including the release of his political opponent Yulia Tymoshenko and the IMF's demand to end to gas subsidies which had functioned as a political tool for Yanukovych[N9]. An alternative to the agreement was swiftly supplied by Russia, which pledged a $15 Billion cash infusion into the Ukrainian economy and an approximately 30% cut in gas prices[N10]. Opposition leaders called for protests, which grew into the so-called "Maidan Revolution" or the "Revolution of Dignity".

Yanukovych's government was replaced with an interim government led by Arseny Yatsenyuk which sought to align Ukraine with the West, leading to increased conflict between Western- and Russia-aligned factions within Ukraine. In response to the change towards a pro-EU government, an uprising in semi-autonomous Crimea installed a pro-Russian leader who called for a referendum on reuniting Crimea with the Russian Federation. The vote was overwhelming in favor of rejoining Russia according to the official tally, although the referendum is seen as illegitimate by the EU and its NATO allies [N11]. The Russian Federation signed a treaty of accession with Crimea on March 18th, 2014, leading to the annexation of the region in the following days.

In response to Russian's annexation of Crimea and ongoing conflicts in eastern Ukraine, Ukraine dropped its position of non-alignment as the interim government announced an intention to join NATO in August of 2014[N12]. Russia expressed opposition to the move, and the conflicts in eastern Ukraine continued. An attempt at a cease-fire was made through the Minsk Agreements in 2015, but this agreement failed when both sides accused the other of failing to implement key terms and of violating the cease-fire agreement[N13].

Sweden was still clearly opposed to joining NATO even after the annexation of Crimea. In November 2021, at the congress of the Social Democratic party, Swedish Defense Minister Peter Hultqvist gave a firm statement on his, and his party's, opposition to joining NATO: "The others

know very well where we stand. ... I will definitely never, as long as I am Defense Minister, take part in such a process. That I can guarantee."[N14]

In December 2021, Russia presented a list of security demands to NATO. Among them were demands that the NATO-Russian border be demilitarized and that NATO not admit any new member states, specifically not Ukraine. NATO rejected Russia's demands on principle, indicating that non-members of NATO had no veto over who the alliance would admit [N15].

On February 24th, 2022, the Russian Federation launched a large-scale military invasion of Ukraine. The Swedish government still held to an anti-NATO position. In March, Prime Minister Magdalena Andersson dismissed the NATO question as a "hypothetical" one, saying that Sweden joining NATO under the present situation would be destabilizing to the security situation in Northern Europe.[N16] Instead, Andersson stressed the importance of the EU's collective defense clause, which states: "if an EU country is the victim of armed aggression on its territory, the other EU countries have an obligation to aid and assist it by all means in their power." [9]

A growing discussion in Sweden about possible NATO membership began, culminating with a joint security meeting between Swedish and Finnish Prime Ministers Magdalena Andersson and Sanna Marin on April 13th, 2022. At the press conference, both PMs indicated that the question of whether both nations would join NATO or not would be decided in "weeks, not months" [N11]. By March, polls began to show a majority of Swedes (51%) in favor of joining NATO, with only 27% opposed[N17].

In May, after Sweden had made its desire to join NATO clear, Recep Tayyip Erdoğan, President of the Republic of Turkey, indicated that Turkey was opposed to Sweden joining NATO due to a perceived support for Kurdish terrorist groups by Sweden [N18].

Nevertheless, on May 16th, 2022, Sweden and Finland both formally applied to join the alliance after several months of public speculation. Neither country held any public referenda on the matter, a point raise in opposition by some.

No major public debate followed, however adjacent political questions emerged. Notably, on June 7th, Sweden's Minister of Justice, Morgan Johansson, survived a No Confidence vote in the Riksdag thanks to the support of a Kurdish-Iranian Swedish MP, Amineh Kakabaveh. In exchange for supporting the Social Democrat-led government during the vote, Kakabaveh demanded that the government would "support the Kurds" and that Sweden would affirm that "people from those organisations coming to Sweden are not terrorists," a position in direct opposition to Turkish demands which had the potential to jeopardize Sweden's NATO ambitions [N19].

On September 11th, 2022, a general election was held in Sweden in which the Social Democrats, long opposed to joining NATO, lost their ruling majority to a coalition led by the Moderate Party. The topic of Sweden joining NATO was not a major issue in the election campaigns which instead focused on rising crime and cost of living.

Generally, arguments against NATO consist of appeals to the diplomatic benefit of a neutral position, complaints due to a lack of a public referendum, opposition to increased military spending, low opinion of the foreign policy of other NATO allies (especially Turkey and the US), and fear of making Sweden a target for Russia. Pro-NATO arguments tend to focus on the vulnerability of Sweden to any Russian aggression, desire for greater international cooperation, and support for Finland and Ukraine.

1.3.4 Current Status

As of the submission of this report in March of 2023, 28 of the 30 NATO members required to ratify the applications for membership from Sweden and Finland have done so, but Turkey has still not voted to accept Sweden as a member of NATO. Sweden has accused Turkey of expanding

their demands, increasing the demanded extraditions to 130 individuals Erdoğan accuses of terrorism[N20], including journalists whom Ankara only officially accuses of ”insulting the President of Turkey” [N21]. The Swedish Supreme Court holds the final decision over which extraditions are constitutional and has not always ruled in Turkey’s favor[N22].

2

Theory and Related Work

This chapter explains the theory behind the methods explored in this paper. This background includes previous research into echo chambers, an explanation of the NLP models that are used in this paper, and a basic introduction into graph theory and algorithms for community detection. As this paper is only a study of methods and applications, none of the models and algorithms described in this chapter were developed by the author of this paper.

2.1 Echo Chambers and Social Media

Previously, it was stated that the underlying hypothesis set to be examined by this work is that language use among social media users will shift as a result of the different opinions held by users within different groups. This divergence of opinion between groups is an expected outcome of the formation of echo chambers. There exists several explanations for the formation of echo chambers, as well as evidence that communities do tend to become more polarized with time.

”Confirmation Bias” is the phenomenon by which people tend to put more faith in information which agrees with their beliefs. Confirmation bias represents itself in three main ways: bias in information seeking, bias in information interpretation, and bias in information retention. Examining echo chambers in social media generally means focusing on the first of those three.

One of the potential driving forces behind social media echo chambers is the idea of ”selective exposure,” a concept where users of social media are free to seek out content that they wish to engage with[10]. When users are free to seek out content, they tend to seek out content generated by users with shared values, interests, or opinions[11]. This can lead to the creation of an echo chamber, as users interact primarily with other users of similar mindset.

Once an echo chamber has formed, there is a tendency of that group of individuals to become more polarized in their opinions. There are various theories for the mechanisms of this, from social comparison to the effects of repetition of persuasive arguments[12]. Additionally, it has been shown that reading other user-generated content can lead to the reader’s opinion being affected even if that reader is not actively engaging with the content by replying[13].

There are many examples of ways in which polarized groups can use language differently within their echo chamber. An extreme example is the adoption of code words by fringe groups, so-called ”dog whistles” which are innocuous words with a double-meaning. Research into dog whistles among Far-Right groups and terrorist organizations has shown that the use of coded terminology behaves very similarly to any other language and that this new ”language” can be acquired through social media[4]. A less extreme example is that consumers of different media sources can have measurably different understanding of the same word as a result of their different exposure experiences[14]. The hypothesis is that users with these different understandings of meaning would therefore tend to use the word in measurably different ways.

This tendency to assign changed meaning to a single lexical term is called ”semantic shift” [15]. Semantic shift is the primary method through which language change is measured in this work. It is

a relative measure, as the arbitrary nature of language prevents there being any absolute ”meaning” to compare any word use to.

It needs to be stated that arguments against the existence of social media echo chambers exist. These arguments make strong cases that differences in information seeking do not alone account for the rise of political polarization in social media, but also that differences in information interpretation play a large role. Research on the political leanings of American Twitter users has shown that exposure to the opinions of political opponents can actually have an increasing effect on polarization[16]. Similar studies have found that, while repeated exposure to partisan media can reinforce opinions expressed by that media, the relationship to out-group media is not so simple as more diverse media consumption leading to less polarization[17]. Alternative causal explanations for polarization in social media other than echo chambers have been proposed, and social network modeling and simulation have yielded some evidence that the selective exposure hypothesis is not the most accurate explanation of the polarization phenomenon[18].

2.2 Word Embeddings

In order to perform any meaningful computation on a set of words, the words must be in a form that a computer can process, which is generally numeric. A word embedding is a numerical representation of a word.

The models used in this paper use a vector word embedding. This means that each word is represented by a vector of some dimension, n. Any meaningful relationship between words should therefore be expressed as a relationship between the vectors which represent those words.

Figure 2.1 shows a very typical example. Here, the Swedish word *kung* for ”king”, and *drottning* for ”queen” are related in the same way as the vector for *man* for ”man” and *kvinna* for ”woman”. A well-trained word embeddings would show a relationship between the vectors representing these words such that $kung - man + kvinna = drottning$. Note that word embeddings are usually vectors in 50-300 dimensions, much more than the 2 shown here.

Similarity in word vectors does not necessarily mean the words have the sentiment. Terms such as *dyr* for ”expensive” and *billig* for ”cheap” may have word vectors that are very closely related despite a human reader interpreting them as direct opposites because these are both adjectives which can be used to describe very similar things in the same contexts.

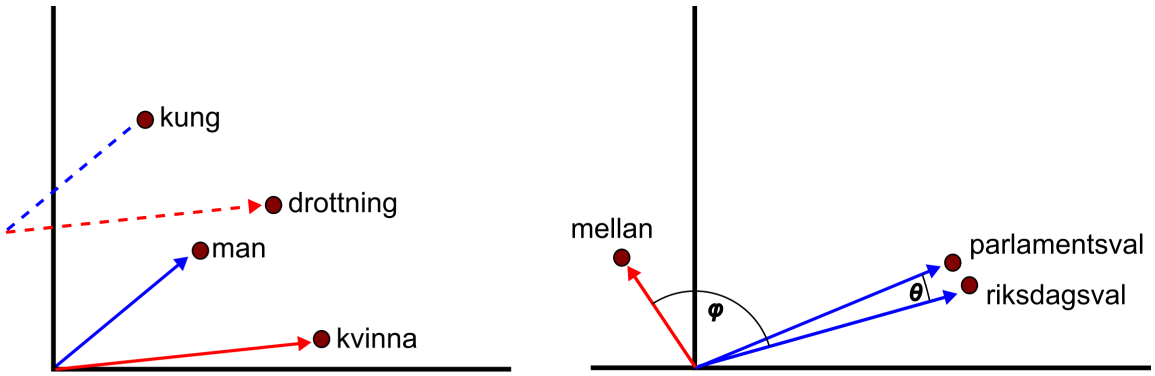

Figure 2.1: (Left) A typical example of word vectorization. (Right) An illustration of similar terms with high cosine similarity as compared to an unrelated term with low cosine similarity.

It is these embeddings which are expected to show semantic changes in words. If the vectors derived from a language corpus shift in meaningful ways and point in new directions relative to

each other, that can be taken as evidence that the use or meaning of the words represented by those vectors has changed.

Prior research supports this assumption. Comparison of word embedding vectors has been used before by a researchers from Georgetown University who were trying to detect novel drug terms to successfully detect new uses of existing language tokens being spread in a community[19], suggesting the possibility that novel uses of terms in other contexts can also be identified. This research was conducted on a non-polarized subject and did not need to consider different opinions or divergent meanings of the novel terms between different sides of the issue, which represents a new and unique challenge.

A common way to make this comparison is cosine similarity, which can be used to identify semantic variation in common words. Two word embedding vectors which refer to words used in the same context will point in similar directions and thus have a higher cosine similarity. In Figure 2.1, vectors representing similar words referring to elections, *parlamentsval* and *riksdagsval*, have a small difference in their angle, θ, giving them a high cosine similarity, while an unrelated word *mellan* is separated by a much larger angle, φ, leading to a lower cosine similarity. Cosine similarity falls in the range [-1, 1] and is defined as:

$$\text{cos sim}\,(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\|\|v_2\|} \tag{2.1}$$

Cosine similarity as a metric to detect changes in vector word embeddings has been used in a similar case to study semantic shift in social networks, where the vectors considered v_1, v_2 representing the vectors for the same word embedded within a community at two different times [20]. That work focused on the stability of terms as a function of the structure of predefined communities and did not assess the semantic change of a given token across different communities.

A study of word embedding models related to Swedish parliamentary motions was conducted in 2021[21] and serves as a foundation for some of the methods in this work. Word embeddings were examined to determine the meaning and attention that different parties gave to different terms. This work expands upon that study by attempting to apply the methodology to social media and attempting to define the user community from the data structure, rather than using predetermined membership that the political party of the motion authors gave.

2.2.1 Word2Vec

There are several models which can produce these vector word embeddings. The one used in this project is a model developed by Google in 2013[22] called ”Word2Vec”. The Word2Vec model was introduced with two different model architectures: The Continuous Bag of Words (CBOW) and skip-gram (SG) model. Both are shallow neural networks that use supervised learning to find a set of word vectors representing the vocabulary of an input text. The difference between the two models is that CBOW attempts to find a word based on a context, and skip-gram attempts to find a context based on a word.

2.2.1.1 Continuous Back of Words

In order for supervised learning to work, there needs to be a labeled set of training data. The Word2Vec algorithm self-generates this training data from the text by inventing a puzzle for the network to solve. Given a context window of m size, the CBOW model uses m prior and m later words to predict a missing word based on a learned context.

Consider a text input that says *nato är en militär allians mellan nordamerika och europa*, translating to ”NATO is a military alliance between North America and Europe”. CBOW can invent a

puzzle by omitting a word, in this example *nordamerika*.

Figure 2.2 illustrates this process. First, each word is reduce to a one-hot vectorization. This necessitates that the length of each input vector is v, the total number of unique tokens in the vocabulary, which is 9 in this problem.

The words in the context window, m, are passed through the trainable matrix $\boldsymbol{W}$, which is $v \times n$, with n again being the vector dimension. As each word is a one-hot vector representation in this step, the contribution from each word to each hidden neuron will only be a single matrix row. These values are either summed or the mean value is taken (depending on network settings) to fill a hidden layer of length n, the length of the word vector. This layer passes through an output matrix, $\boldsymbol{Z}$ of size $n \times v$ and a softmax activation function to produce an output vector, $\boldsymbol{y}$. The loss function is found by comparing the output, $\boldsymbol{y}$ to a target vector, $\boldsymbol{t}$, which is the one-hot encoded missing word, which is $nordamerika$ in this example. Back-propagation is used to update the matrices $\boldsymbol{Z}$ and $\boldsymbol{W}$ to correct any errors.

This process is repeated for all of the sentences in the input text, with m varying up to some determined maximum size. The usable word vectors are taken from the trained matrix $\boldsymbol{W}$, with each row being an n-dimensional vector representing the word corresponding to that row.

CBOW is considered to be better at placing very common words than the skip-gram model[23], which will allow a more accurate context for the most common tokens of interest in the dataset.

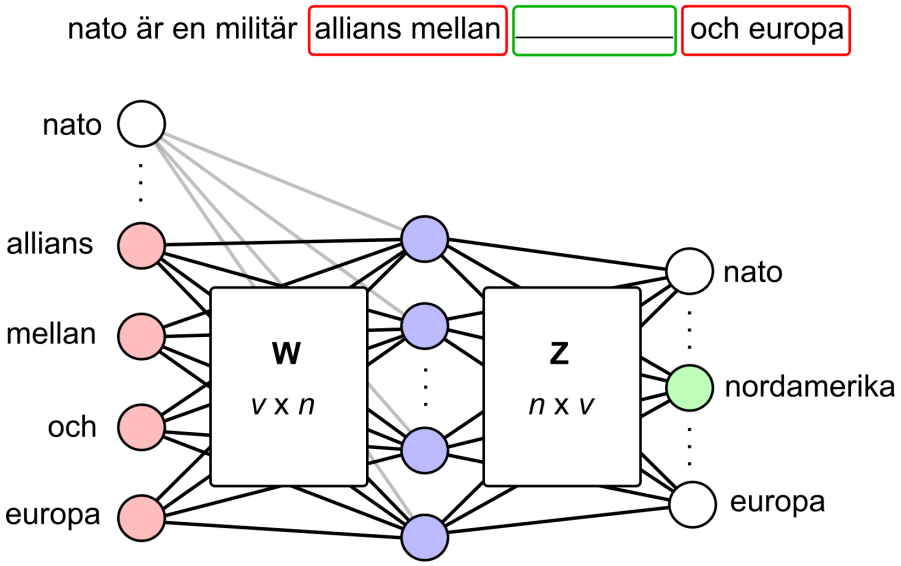

Figure 2.2: The CBOW Model for Word2Vec training on the input sentence *nato är en militär allians mellan nordamerika och europa*. The mean of input vectors of words *allians, mellan, och*, and *europa* are passed through the weight matrix $\boldsymbol{W}$ to a hidden layer of dimension n. The output is passed through the weight matrix $\boldsymbol{Z}$ with a softmax activation function, given the predicted vector amongst the input vocabulary.

2.2.1.2 Skip-Gram

The skip-gram model takes a single word as input, and runs a similar process to attempt to guess the context. In this example, the input sentence *nato arbetar för internationellt militärt samarbete*, or ”NATO works towards international military cooperation” is given to the model with a window size of $m = 1$.

The model takes the word *militärt* and attempts to estimate the context. There is only a single input vector representing the word of interest, *militärt*, which is multiplied by the W matrix, again utilizing a single row due to the one-hot vectorization of the input.

Figure 2.3 shows the process for this model. For each word that must be guessed, a separate multiplication is done between the hidden row and the output matrix, $\boldsymbol{Z}$. Again a softmax activation function is used. In this example, the first context word to be guessed is *internationellt*, and the second is *samarbete*. Failure to estimate the correct word results in a loss, which is used to adjust $\boldsymbol{Z}$ and $\boldsymbol{W}$ through back-propagation.

The computational complexity of the skip-gram model is higher that CBOW[24], requiring more training time. However, the skip-gram model tends to perform better at finding a placement for less frequent words, as it must train a word/context matching for any words above the minimum frequency threshold. Skip-gram performs worse on common words than CBOW, as a very common word becomes a likely guess for most other words' contexts.

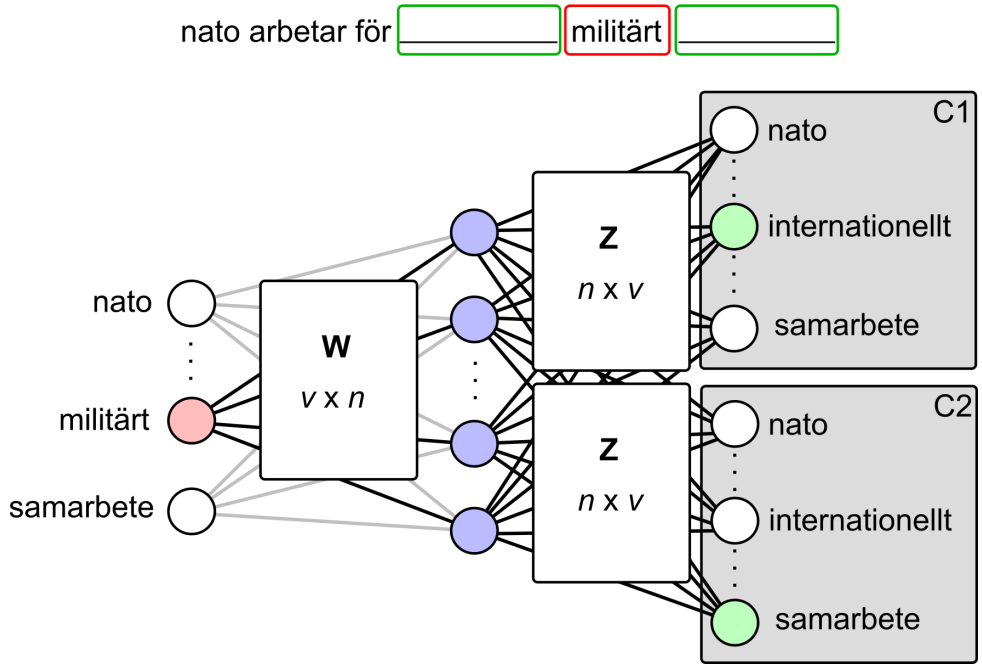

Figure 2.3: The Skip Gram Model for Word2Vec training on the input sentence *nato arbetar för internationellt militärt samarbete*. The weights in $\boldsymbol{W}$ for the one-hot vector representing the input words are applied to each hidden neuron, and then multiplied by the output matrix $\boldsymbol{Z}$ for each context word that should be guess.

2.3 Community Detection

2.3.1 Graphs

To perform analysis of networks, it is common to format the data into a graph. A graph is a structured set of objects which represents how groups of objects are related. More formally, it is a set of sets of elements called *nodes* or *vertices*. A graph limits each set of elements to two members and calls the set an *edge*. A generalized type of graph called a hypergraph exists which does not put such a restriction on the sets; a *hyperedge* can join any number of elements. But in this work, only the typical graph where each edge connects two vertices is used.

A graph, G, is defined as $G = (V, E)$, where V is the set of vertices and E is the set of paired vertices. A graph's *order* is the number of vertices, and its *size* is the number of edges. For each vertex, the number of edges that it is a member of defines its *degree*. A vertex that has edges towards four other vertices would have a degree of four. The degree of vertex i is written as d_i. A vertex with no connections can still be part of the graph–that vertex has a degree of zero. Figure 2.4 shows a simple graph.

A simple graph is the most common type of graph. Here, every pair of vertices can only be joined by one edge. In this case, the maximum possible degree of a vertex in a graph of order n is $n - 1$, the case when the vertex shares an edge will all other vertices. If self-edges (a vertex connected to itself) are allowed, which is sometimes the case in graphs, then the vertex shares an edge with itself and the degree of the vertex is n.

A typical data structure used to store a graph is an *adjacency matrix*. The adjacency matrix of an order n graph is an $n \times n$ matrix $\boldsymbol{A}$ whose elements a_{ij} are the number of connections between vertex i and vertex j. In a simple graph, $a_{ij} \in \{0, 1\}$. In a multi-graph, which allows multiple edges between the same two vertices, this value can be any integer. For the types of graphs discussed so far, $a_{ij} = a_{ji}$, i.e. $\boldsymbol{A}$ is symmetric.

The vertices and edges of a graph can have properties assigned to them based on the purpose of the analysis. If the vertices were cities, then the vertices may be assigned values such as population or priority, while the edges could represent distances. This kind of graph might feature in a shortest-path problem. When edges have numeric values assigned to them in this way, this value is called the edge's *weight*. When a graph is weighted, then the adjacency matrix A contains the weight of the edge between the two vertices. A vertex's degree in a weighted graph is the sum of all weights of the edges of which it is an element.

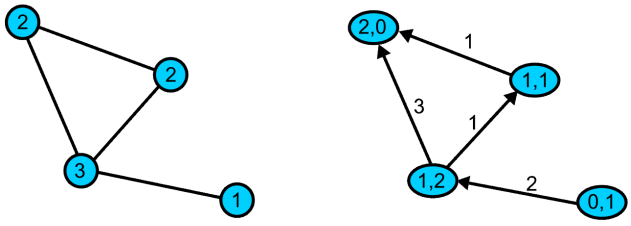

Figure 2.4: (Left) A simple graph. The vertices are labeled with their degree. (Right) A directed graph, with the vertices labeled with the in degree and out degree. The edges have weights, a setup for a trivial shortest-pah problem.

2.3.2 Directed Graphs

In some contexts, it is necessary to consider that the connection between two elements is not bidirectional. For example, if the vertices represent states in a process, it may not be possible to undo certain operations which move the process from one state to another. These types of graphs are called directed graphs. Figure 2.4 shows an example of a directed graph.

In a directed graph $G = (V, E)$, each edge $e_{ij} \in E$ is an ordered pair of vertices representing a one-directional edge going from vertex i to vertex j. Because the number of edges starting from a vertex and arriving at the vertex are not necessarily the same, there now exist two different properties of the vertex: *in degree* for the number of edges starting from the vertex, and *out degree* for the number of edges ending at the vertex. These are represented by d_i^+ and d_i^- respectively. In a

weighted directed graph, these values would be the sum of the weights of the incoming or outgoing edges.

In a directed graph, the adjacency matrix $\boldsymbol{A}$ is no longer symmetric, as $a_{ij} \neq a_{ji}$. In a weighted directed graph, such as the one shown in Figure 2.4, a_{ij} is equal to the weight of the edge leaving vertex i towards vertex j.

2.3.3 Community Detection Algorithms

In large networks, it is common for small subsets of the network to be more closely connected to each other than they are to the larger network. These subsets are called *communities*. The community structure of a network is of considerable interest to social science researchers; even finding the existence of communities within a network can provide useful information about the behaviors of the network[25]. In terms of network topology, communities are vertices which have dense connections between each other, but sparse connections to vertices outside of the community.

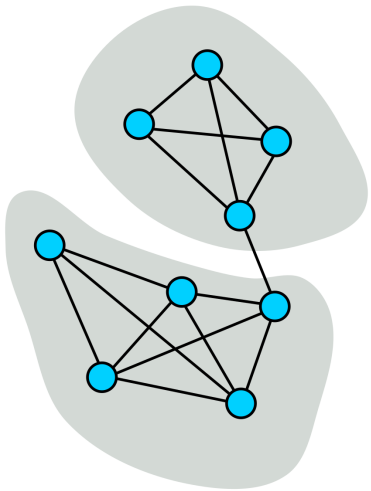

Figure 2.5: A graph showing clear community structure.

The echo chamber hypothesis expressed in the previous sections translates to community structure in network graphs in that we suppose that vertices in well-defined communities within the network share high similarity in certain traits with other vertices in their community, but low similarity with vertices outside of their community.

Figure 2.5 shows a graph with an obvious community structure; the four top vertices are densely connected to each other, while the lower five vertices are densely connected to themselves and only sparsely connected to the other four. Detecting communities in large graphs is an ongoing challenge; in a large graph, with tens of thousands or even millions of vertices, it is not possible to find such clear and obvious partitioning as in this example.

Several algorithms have been developed towards this end. Those considered in this research are the Directed Louvain and Leiden algorithms. These algorithms have been shown to perform well in finding well-connected graphs among Twitter users [26], so those clustering algorithms are employed. The same research found similarly good results for the Infomap algorithm as well, however that algorithm showed a tendency to return larger super-communities which is not expected to be a suitable approach for finding opposing sides in a political debate.

2.3.3.1 Louvain Algorithm

One of the most common metrics for evaluating a partition in a graph is *modularity*. Modularity is the fraction of edges between vertices in a community minus the expected value of a random distribution of edges between those same vertices[27]. This modularity uses a model of a random graph where the vertices retain their degree to determine the expected fraction of edges. In such a random graph, the chance of these being an edge between two nodes is $\frac{d_i d_j}{2m}$, where m is the size of the graph (number of edges). The modularity is evaluated based on a partitioning of every vertex i into some community c_i. Thus, the modularity can be expressed as:

$$Q = \frac{1}{2m} \sum_{ij} \left[\boldsymbol{A}_{ij} - \frac{d_i d_j}{2m} \right] \delta_{c_i, c_j} \tag{2.2}$$

A maximum value for Q represents the optimum partitioning using modularity as the evaluation metric. Note that there are some limitations to using modularity as an evaluation metric which are discussed in a later section.

An algorithm was developed which maximizes this metric in an efficient manner. The algorithm is commonly called the Louvain Algorithm[28] after the researchers from the University of Louvain who created it.

The Louvain Algorithm initializes the vertices as members of their own individual communities. From there it attempts to maximize the modularity in iterative steps of two phases.

In the first phase, for each vertex i, the neighboring vertices j are considered. The gain in modularity from moving vertex i to the same community as some neighbor j is evaluated, and the move which results in the maximum gain in modularity is performed. If multiple moves result in the same gain, then the move chosen is randomly decided. This phase ends when a local maximum is found.

The second phase creates a new graph whose vertices are the communities found in the first phase. The weights between two vertices are the sum of the weights of the edges between vertices in each of the two communities represented by the two vertices. Connections within the communities are represented as self-loops.

From here, the first phase can be repeated. Figure 2.6 illustrates the two phases of the algorithm over several steps.

This algorithm applies only to undirected graphs. A variation for directed graphs has been developed[29] using a modularity measurement based on in degree and out degree[30]. Also, the formulation for the modularity in a weighted graph must take into account that the expected number of connections is not a function of total number of edges, m, but rather a function of the sum of all degrees, w. This gives the final formulation for the modularity used:

$$Q_d = \frac{1}{w} \sum_{ij} \left[\boldsymbol{A}_{ij} - \frac{d_i^+ d_j^-}{w} \right] \delta_{c_i, c_j} \tag{2.3}$$

2.3.3.2 Leiden Algorithm

One risk of the Lovain Algorithm is the creation of unconnected communities. Figure 2.7 shows an example of how the greedy Louvain algorithm can cause a break in a community by moving a connecting vertex into another community to increase modularity. A proposed solution in the form of the Leiden Algorithm was given by researchers at the University of Leiden who identified this problem[31].

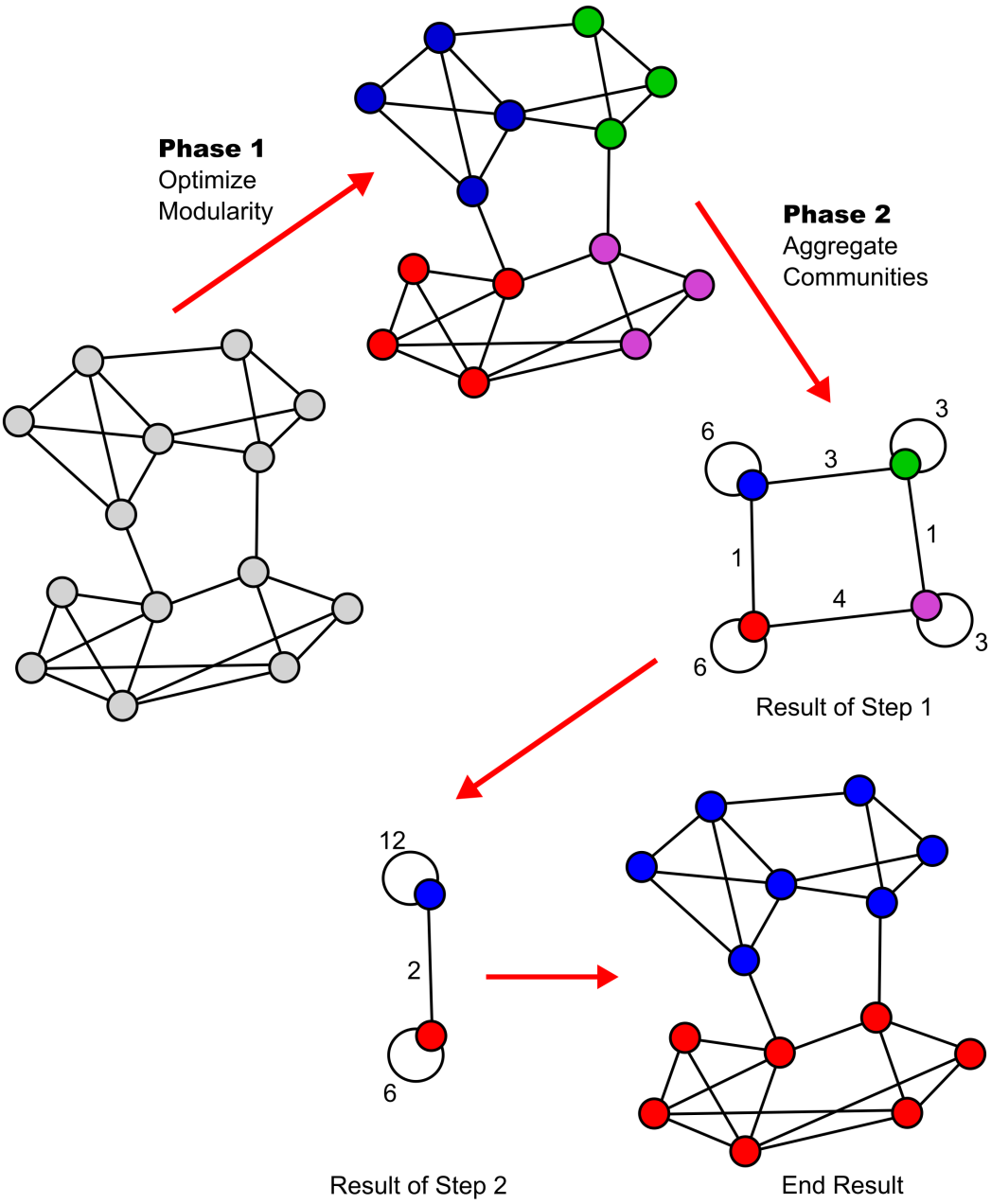

Figure 2.6: The Louvain Algorithm's two phases. In the first step, the first phase is able to assign the vertices into four communities before reaching a maximum modularity. In the second phase, a new graph is created with each of those four communities as a vertex. These two phases repeated in another step, leading to a final partitioning.

The Leiden Algorithm differs from the Louvain Algorithm in few key ways. First, the Leiden Algorithm assigns vertices to a new community according to a probability based on the size of the potential increase in the objective function, rather than choosing the largest gain as in the Louvain Algorithm. More importantly, the Leiden Algorithm has a refinement step before creating the aggregate communities in the second phase. This refinement sets all vertices within a community into their own partition and repeats the moving process, restricting vertices to forming communities with those that they shared a partion with in the initial phase.

The newly defined partitions are used as the vertices in the new aggregate network, however they

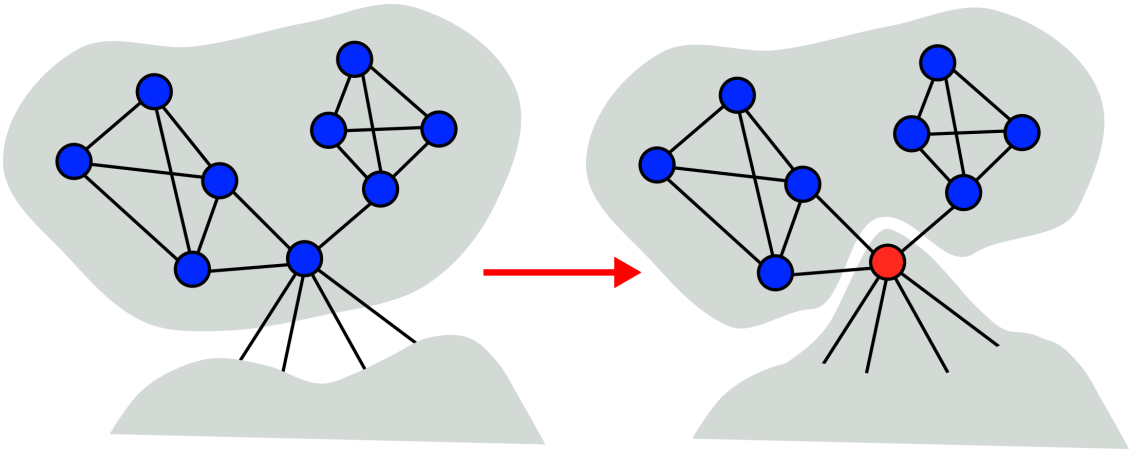

Figure 2.7: The Louvain Algorithm may assign a node to a community in what is a locally-optimal move, but that disconnects communities from each other.

are initially assigned partitions according to the pre-refinement partitioning, as in Figure 2.8. This guarantees connectedness[31].

2.3.3.3 Constant Potts Model

Modularity-based community detection methods such as the Leiden Algorithm operate on an intrinsic scale set by the graph size. This leads to a problem known as the ”resolution limit”, whereby modularity-based community detection algorithms cannot find small structures inside of larger communities[32]. The modularity equation in 2.3 includes a term for the total number of edges in the graph, m, used to estimate the number of expected edges between two nodes. When summed over entire communities, this penalized communities below a certain threshold size, leading the algorithm to merge these communities into larger ones.

Another quality function than modularity has been proposed which introduces a scaling factor to eliminate the resolution limit problem[33]. This method is called the Constant Potts Model.

$$Q_d = \sum_{ij} \left[\boldsymbol{A}_{ij} - \gamma \right] \delta_{c_i,c_j} \tag{2.4}$$

This objective function can be used instead of modularity in the Leiden Algorithm. Here, the parameter γ replaces the expression for the expected number of connections in (2.3), allowing a tunable search for communities of arbitrary size as this model will break a single community into two separate communities if the link density between them is lower than the resolution parameter.

This approach may be of considerable benefit in a large social network which contains numerous smaller communities inside a few larger ones.

2.3.4 InfoMap Algorithm

Another common algorithm for community detection is the Infomap Algorithm. This algorithm is based on the Map Equation devised by Rosvall and Axelsson[34] and works to minimize the length of a code L that would be needed to describe precisely the trace of a random walker moving across the graph by assigning a unique and prefix-free code to each node in the network, i.e. no node has a code which is a prefix of another node’s code, so that the data can be streamed without breaks. The Infomap Algorithm was developed using the Huffman Code method [35] which achieves this in a minimal form.

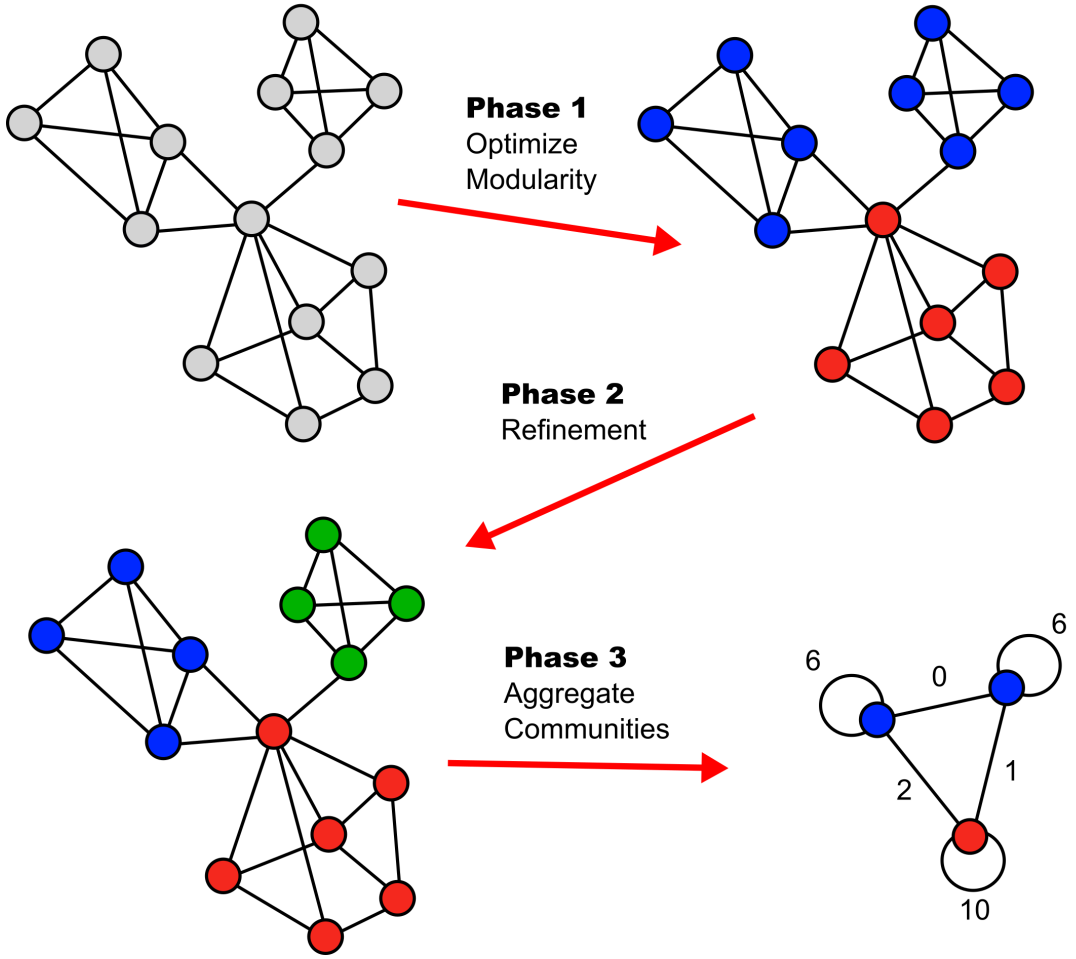

Figure 2.8: The Leiden Algorithm's three phases. In the first step, the first phase is able to assign the vertices into two communities before reaching a maximum modularity, which can result in the split community as shown if the randomized assignment is unfortunate. In the second phase, the refinement prevents these unconnected communities from being combined in the aggregate phase. The third phase sets up an aggregation. The refinement of the next step would prevent the two blue nodes (representing the blue and green communities) from being combined as there are no mutual connections.

Even with a minimal form, it is clear that the code for a single node can become long as the number of nodes increases. The solution proposed with the Infomap Algorithm is to develop a partitioning, M, of the graph where nodes are grouped into modules. Each module has its own codebook, so the code for a node only needs to be unique within its module. An additional codebook, the index codebook, is added with the entry codes indicating which codebook should be used for decoding the subsequent nodes. An illustration of this is given in Figure 2.9.

This is achieved by running a similar process to that of the Leiden Algorithm, but rather than maximizing modularity, the goal is to minimize the Map Equation:

$$L(M) = q_{\curvearrowright} H(Q) + \sum_{i=1}^{m} p^i_{\circlearrowright} H\left(\mathcal{P}^i\right) \tag{2.5}$$

Here $H(Q)$ is the frequency-weighted average length of codewords in the index codebook, $H(\mathcal{P}^i)$ is the frequency-weighted average length of codewords in the codebook i, $q_{\curvearrowright}$ is the probability to exit a module, and $p^i_{\circlearrowright}$ the probability to stay within the module i. The exit probabilities

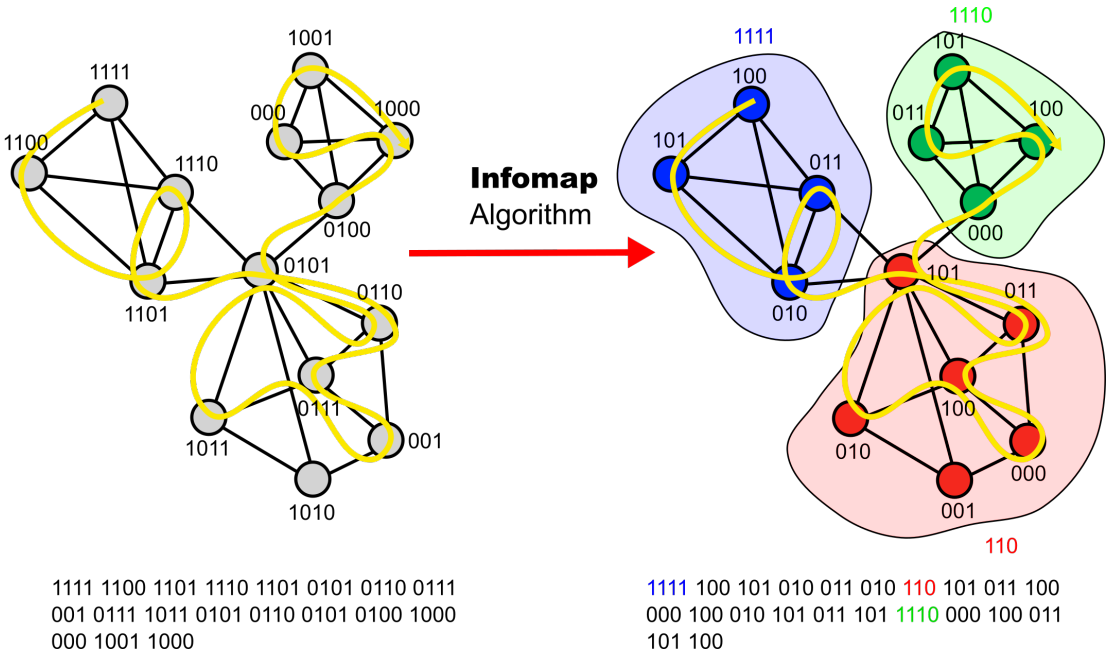

Figure 2.9: The Infomap Algorithm and the resulting reducting in code length for a sample random walk. Here, recoding the nodes into three separate codebooks and using an index codebook for switching between modules reduces the total code length of this random walk by about 9%, but in larger, more complex graphs the reduction is greater.

are simply taken from the relative weighting of the edges from a node to nodes in its module and to nodes outsides its module. The algorithm forms modules out of regions where the random walker has a high likelihood to stay in for a long time. This causes the Infomap Algorithm to naturally assume a structure based on the flow within the graph and the dynamics of the network, which is of considerable importance when the hypotheses are based on the flow of information.

3

Methods

This chapter details the methods performed as part of this research. The process of data gathering is explained, followed by sections about how the data was grouped by time period and by user using the community detection algorithms discussed in the previous chapter. A strategy for estimating user opinion via manual annotation is then outlined.

The second half of this chapter covers the use of Word2Vec models. The text pre-processing pipeline for the dataset is explained, followed by the process setting up pre-trained models language models to use as a base for fine-tuning using the text in the dataset. The chapter concludes with details of the setup for training Word2Vec models on the dataset and the selection process for those words whose vector embedding will be examined in the following chapter.

3.1 Data Gathering

The social media data for this project was taken from the micro-blogging platform Twitter. Twitter users post ”tweets” which are limited to 280 characters in length. This keeps each item in the dataset at a relative similar length. The user base for Twitter is very large and touches all segments of society, which helps to reduce concerns about bias in the dataset that would arise if the source material were from a service used only by individuals of a given education level or political leaning. Many government ministries and members of the Swedish Riksdag have official Twitter accounts, as do media sources, so the postings on Twitter are expected to be representative of the broader media coverage of a topic and contain responses by the average citizen to official statements from authoritative sources.

The data gathered consists of the text of the tweets, as well as the reactions to those tweets by other users. Users can ”like” a tweet, indicating approval, or they can ”retweet” or ”quote” the tweet in their own postings, and they can reply directly to the original poster. Other users can see the replies to these posts and engage other users in conversation.

In addition to tweets themselves, data about the users who posted the tweets was also collected. Users can ”follow” other users, meaning that when they log into Twitter, they will see recent posts by users of their choice on their home page. Users can also create or join lists; those who follow a list will see posts by members of the list on their home page. This allowed the creation of the user network graph.

The total size of the dataset is 1 188 556 tweets, made by a total of 64 315 users participating in 507 359 conversations. Of these, 329 336 are retweets–just a reference to another user’s posting. This leaves 859 220 original tweets.

3.1.1 Search Criteria

A year’s worth of tweets was considered, with a starting date of September 11th, 2021 and an end date of September 11th, 2022, the date of the Swedish national election. This was chosen so

that there would exist a ”baseline” NATO perception in the data before the events surrounding Russia’s invasion of Ukraine and the discussion of Sweden’s entry into NATO. When this project was planned, it was expected that Sweden’s entry into NATO would be an important election topic. That was later found to not be the case.

Because the topic was Sweden’s entry into NATO, only Swedish posts are considered. While there is a lively debate about NATO expansion among English-speaking Twitter users, collecting this data would lead to problems with geographical variation when considering language user patterns as well as too many results outside the scope of the original context. Collecting only Swedish data eliminates these concerns. Furthermore, it is not practical to run language analysis tasks on a dataset containing multiple languages.

Tweets were collected if they contained any one of a set of search terms related to NATO. The keywords consisted of NATO-related terms in Table 3.1, alliance- and neutrality-related terms in Table 3.2, and terms expected to detect related discussions in Table 3.3. The terms in the third table are less strictly related to NATO than the others, but during the time period we are interested in they are seldom used in any other context.

#NATO	#NejTillNATO	#JaTillNATO	#WeAreNATO
#NATODebatt	#NATOAnsökan	#NATOFrågan	#NATOSummit
#NorgeUtNATO	#SverigeNATO	#NATONu	#NoToNATO
#SwedenNATO	#NATOSverige	#NATOtroll	
nato	natomedlemskap	natomedlem	natomedlemmar
natoval	natofrågan	natomötet	natoland
natoländer	natolandet	natos	nato:s
natoansökan	natodebatt	natodebatten	natomoståndare
natosoldater	natoallierad	natosamtal	natoavtal
natoavtalet	natoanslutning	natoanslutningen	natoprocess
natoprocessen	natotroll	natotroller	natobesked
artikel 5	csto		

Table 3.1: NATO-related search terms.

#alliansfrihet	#alliansfri	#alliansfria	#alliansfritt
alliansfriheten	alliansfrihet	alliansfri	alliansfritt
alliansfria	allianslös	neutralitet	neutraliteten
självständig utrikespolitik	säkerhetsgaranti	försvarsallians	militärallians
militär allians	suveränitet	allians med turkiet	försvarssamarbete
angreppspakt	försvarspakt	försvarspakten	försvarsalliansen

Table 3.2: Alliance-related search terms.

#folkochfred	#imperialism	dö för turkiet	finlands sak är vår
diktera vår säkerhetspolitik	erdogan diktera		

Table 3.3: Additional search terms tangentially-related to the NATO conversation.

The Twitter keyword search is not case-sensitive. Variations of these terms which includes a hyphenated form, such as *Nato-länder*, were also included due to the lack of consistency in spelling on Twitter. Some combinations and forms do not appear in the list if a search found that these terms

never appeared without another of the search terms in the tweet. Also included in the search are any tweets which twitter has flagged by Twitter with a context ID as being related to NATO.

These keywords were chosen to keep the dataset relevant specifically to the question of Sweden applying to join NATO. Even though the context of the NATO debate included other world events– such as Russian's invasion of Ukraine and Turkey's conflict with the PKK and YPG–including broader terms into the dataset would have expanded the scope of the work beyond its initial intent of analyzing how different sides of a debate use language. Terms such as *Erdogan diktera* capture complaints that many Swedes had about a perceived acquiescence to Turkish demands by the government in exchange for support in Sweden NATO bid.

There were many terms which could indicate a NATO-related discussion but were not unique to the NATO topic. For example, the term *allians* often was used to refer to NATO, however it is used more commonly to refer to the Center-Right party bloc in Sweden's Parliament. Indeed, before 2022 the hashtag #alliansfritt was used primarily to express support for a Left- or Center-Left government which would keep *alliansen* out of power. It is only during the period of Sweden's NATO application that it shows up often enough in NATO-related tweets to include in this search.

Figure 3.1 shows the structure of tweets included in the dataset based on a keyword search. All tweets that are replies to a tweet containing a NATO keyword are included, as are all those which are parents of a NATO keyword tweet, as they provide a context in which another user would have introduced discussion of NATO. Any conversation headed by a NATO tweet is included.

For example, a user tweets a remark saying that they do not wish to send their son to die for Turkey (a possible consequence of Sweden joining NATO, as Sweden would have a mutual defense obligation towards Turkey). This is collected in the dataset due to the keyword "alliansfrihet":

> Alliansfrihet och Parisavtal. Jag vill inte skicka min son att dö för Turkiet och jag vill att vi lever på vad jorden producerar.

The context of this tweet is found in the parent tweet, which is an open question of why Swedish people might support the Green Party. This tweet is not related to NATO, but it does help to establish the scope of the NATO discussion, as the Green Party are one of the few parties in the upcoming election opposed to NATO membership:

> Hur i helvetet kan folk rösta på Mp vad fan är det för fel på folk i Sverige

Most of the replies to this tweet are not NATO-related. If all tweets made in the same conversation as a NATO-related tweet were gathered, then the dataset would contain almost all political tweets made over the course of the year and greatly exceed the target scope.

3.2 Time Steps

To assess the changes in semantic usage over time, several time steps are considered reflecting milestones in the process of Sweden's application to NATO. The number of tweets included in each time step is given in Figure 3.3. Four time steps were considered which represent four phases of the NATO discussion in Sweden.

Figure 3.2 shows the distribution of tweets which were collected over the year leading up to the Swedish election. The volume of tweets is broken down into original tweets and retweets, which contained no new text.

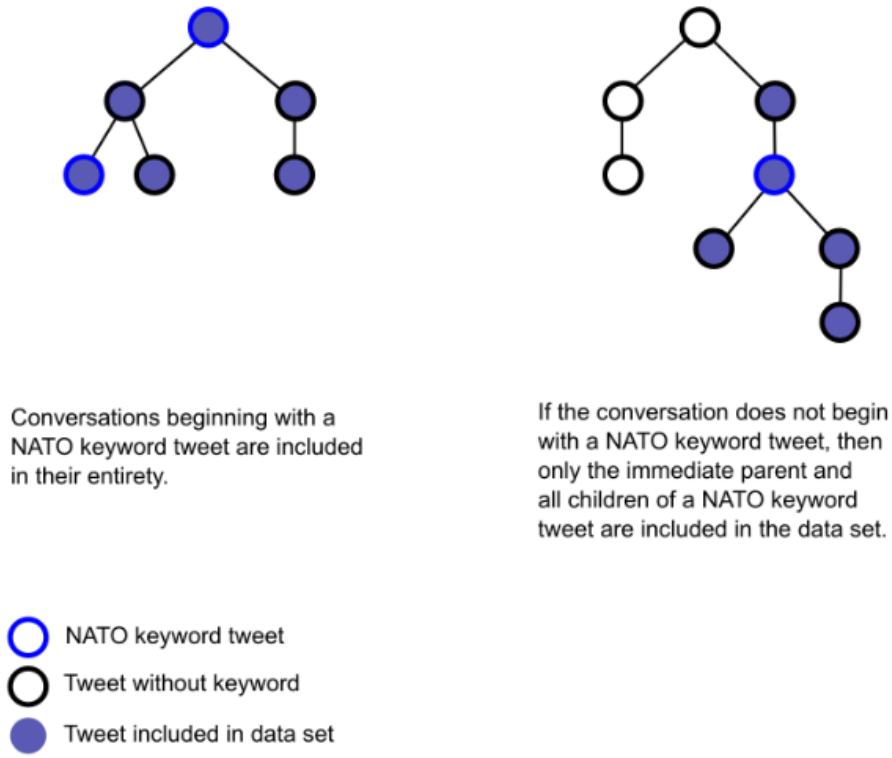

Figure 3.1: Selection methodology for inclusion of tweets in the dataset.

3.2.1 Pre-Invasion

Tweets from 2021-09-11 until 2022-02-24, the date the Russian Federation launched a large-scale military invasion of its neighbor, Ukraine, are considered to be ”pre-invasion” tweets. The volume of tweets about NATO are low, and little is being said about Sweden joining NATO.

There is an increase in activity around mid-January, when the US publicly stated that an invasion of Ukraine is likely [N23].

3.2.2 Post-Invasion

There is a large volume of tweets after 2022-02-24 which reference NATO. Much of the context of the conversation is NATO’s possible involvement in the war and what assistance NATO will provide to Ukraine. Talk of Sweden joining NATO begins here and becomes more common leading up to the cut-off point of 2022-04-13, when a joint press conference was held with Magdalena Andersson, the Prime Minister of Sweden, and Sanna Marin, the Prime Minister of Finland.

3.2.3 Pre-Application

From the date of the press conference, 2022-04-13, until the formal application by Sweden and Finland to join NATO on 2022-05-16, the conversation about NATO in Sweden is heavily focused on Sweden’s NATO accession. It is during this period that Turkish opposition to Sweden’s membership bid emerges. The discussion of NATO membership for Sweden in the final days before the application weighed the consequences of acquiescence to Turkish demands, which some users saw as a compromise to freedom of speech and assembly.

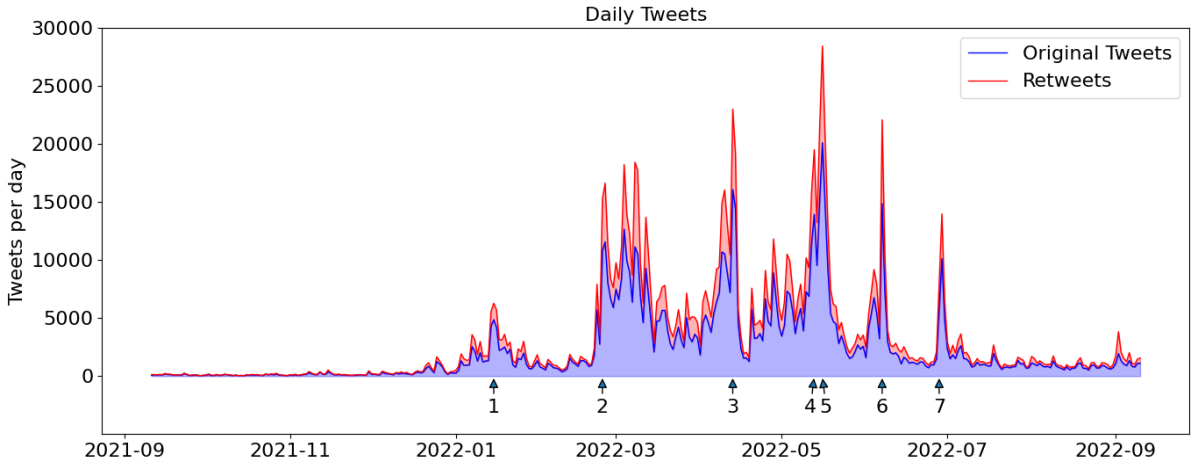

Figure 3.2: Tweets found by the search criteria from 2021-09-11 until 2022-09-11. The notable dates are (1) January 15-19th, Russian military build-up on the Ukranian border, (2) February 24th, the Russian Federation invades Ukraine, (3) April 13th, Swedish and Finnish PMs hold a joint press conference about the decision to join NATO, (4) May 13th, Turkey signals their opposition to Swedish entry into NATO, (5) May 16th, Sweden and Finland formally apply to join NATO, (6) June 7th, ”No Confidence” vote held in the Riksdag for Justice Minister Morgan Johansson, (7) June 28th, NATO Summit in Madrid

3.2.4 Post-Application

After the application to join NATO was submitted on 2022-05-16, there was little public debate in Sweden about joining NATO. The question of joining NATO did not materialize as an election issue. Turkey appeared to drop their opposition after the NATO summit on June 28th, and the NATO question was not a major election issue.

3.3 User Graph

3.3.1 Building a Network

The set of twitter users that have posted tweets in the collected dataset are assembled into a user network graph. This is done so that user communities can be detected using well-known clustering algorithms to allow for the detection of potential echo chambers. Construction of the user network graph did not consider the content of the tweets, as the clustering is meant to reflect only user connectivity by modeling the discussion groups.

The user network graph is setup as a directed graph to show the flow of influence from one user to another. Because the hypothesis is that language that users on Twitter employ will affect other users, the direction of the connections in the graph is always towards the user who is expected to be influenced by another user.

The creation of this graph begins with an empty graph where one node represents each of the user ids associated with authorship of one of the tweets in the tweet data set. Edges are then inserted to model the different types of user connections:

- Edges from a user to all users who have liked one of their tweets
- Edges from a user to all users who have replied to one of their tweets
- Edges from a user to all users who have quoted one of their tweets, including as a retweet
- Edges from a user to all users who are following them

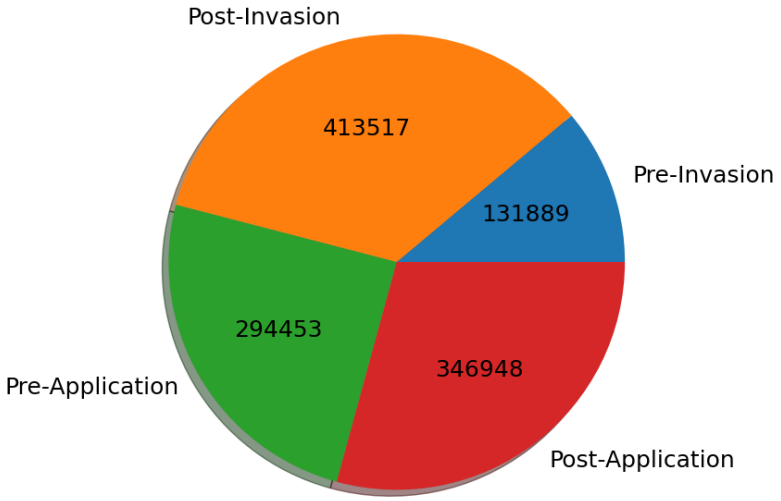

Figure 3.3: Distribution of tweets in the dataset from each of the time steps being considered. The segmentation was done to keep the sizes of the resulting text corpus of each time step as equal as possible, while separating the data based on events and milestones which were likely to have significant effect on public opinion.

- Edges from a user to all users who are following a list that the user is a member of

Twitter users will see a home page showing tweets by users that they are following. If a user looks at their own tweets, they will see replies made by other users, but replies to their own tweets will not appear on the home page unless they are following the user that they have replied to. Replies are not treated as two-directional connections because this behavior of the Twitter platform does not guarantee that a user will see replies to their own tweets, so influence cannot be assumed. For example, the tweet below is unlikely to be read by either NATO Secretary General Jens Stoltenberg or the Turkish President, and even less likely to affect their opinion of the NATO issue.

> @jensstoltenberg @RTErdogan Nej till Nato!!!!

All users with no connections are removed from the graph, and any small groups of users with no connections to the larger graph are also removed. When forming connections between users in this manner, the resulting network had 63546 vertices and 12.7 million edges. Table 3.4 gives a breakdown of the numbers and types of edges. This breakdown is included because the search for a relevant partition was performed in several steps. Most of the user connections were drawn from the network of followers (including followed lists). The overlap between likes and follows was not as large as expected–only about half of the tweets that were ”liked” were liked by a follower of the user.

Because there are multiple types of connections, some which are binary (following) and others which can have multiple edges, different network topographies were considered. Figure 3.4 gives an example of seven users who have connections and the different resulting graphs using the two approaches.

The first network setup used a naive approach to graph topology–any connection was treated as an edge, and all edges were uniformly weighted. This idea was based on the principle that any exposure could lead to an opinion shift and therefore a language shift. Tweets by any user that one followed would appear on one’s home page, so there was an opportunity for influence even if there

Edge Type	Count
Follows	10 334 551
Likes	3 101 154
References	763 366
Replies	382 948
Retweets	209 833
Quotes	22 777

Table 3.4: Breakdown of edges between users.

Setup	Nodes (Users)	Edges	Weighted	Avg. Degree	Median In Degree	Median Out Degree	Tweets
Full Network	63 561	12 760 051	No	201	85	19	845676
Likes Only	53 806	3 101 154	Yes	119	19	1	821103

Table 3.5: Network setups of the user graphs showing the number of users connected by edges of the type in question, the number of edges, and the average degree distributions.

was no direct engagement.

The second network setup considered only ”likes” as a connection between users, and the weight of the edges was defined by the number of likes. This was decided under the presumption that tweets that were engaged with had a greater chance to influence the users. If a user had a strong opinion, they would be more likely to engage, either by retweeting, replying, or liking the tweet. Recent research suggests that, while most Twitter users are not very politically active, those who are tend to favor their side of a political argument with engagements such as likes, quotes, and retweets at a rate of up to 20x times their opponents, while their follows only favor their side around 8x as often[36], suggesting that looking at only these engagements may be a better way to isolate hidden communities. Preliminary testing showed that considering all reactions resulted in low modularity relative to other setups. Furthermore, user engagement via quotes and replies can be misleading– users often quote a political opponent for an opportunity to mock them, and replies are often made to disagree with another user’s posts. Even retweets are no guarantee of support. Given the fact that most of the reaction-based connections were defined by likes, the higher modularity scores of using the likes-only connection, and the greater certainty that a ”like” indicated a user holding a similar opinion to the poster, the quotes, replies, and retweet connections were dropped in this network configuration.

The sizes of the networks created with these two methods is given in Table 3.5. Most users who are members of the likes-only network are represented by incoming connections only (influence from others), so there is a very low median out degree for the user nodes. This is the nature of social media–a few users have very large reach with their content, while most are only consumers whose posts garner very low engagement.

3.3.2 Community Detection

To examine the hypothesis that different user communities will use language in different ways, discrete communities needed to be detected. The community detection algorithms that were used split the entire user network into tightly-related communities so that the text of the tweets posted by users in each of the different communities could be examined as a separate corpus. Four methods were used here: (1) modularity-based community detection with the Leiden Algorithm, (2) a Hierarchical

Leiden Algorithm where communities above a threshold size were broken down further by running the algorithm again on the subgraph represented by the community, (3) the resolution limit-free Constant Potts Model, and (4) the Infomap partition.

3.3.2.1 Consensus Partitioning

All of the algorithms considered are stochastic algorithms; different runs can yield different vertex partitions because of random selection of vertices during the movement phases. It is tempting to run the algorithm several times and simply take the partition with the highest value of modularity achieved, but there may be many partitions with arbitrarily small differences in the modularity, so simply selecting the ”best” partition based on a very small modularity advantage might not give a clear picture of the network structure[37]. For this reason, a consensus clustering algorithm is employed. Each algorithm is run on the data 10 times, and a coassignment matrix D is created for the considered users with elements $d_{i,j}$ equal to the fraction of the runs in which the two users i and j shared the same partition.

Consensus clustering algorithms generally follow a procedure where a original clustering algorithm would be iteratively run on a new graph where the coassignment matrix would function as the adjacency matrix[38], however a dense $N \times N$ matrix would be computationally intractable to process given the number of users in this network. Instead, in the coassignment matrix D only elements $d_{i,j}$ for all neighboring pairs of vertices i and j are computed to keep the coassignment matrix sparse[39].

Edges with weights that were below 0.2 (or only shared a partition with a neighboring vertex in 20% of runs) were set to zero in D, unless that would result in a node having no connections, at which point its highest scoring connection was preserved. The community detection algorithm was run once again on a network with D as the adjacency matrix, and the resulting partition was kept as the final partition for that network setup.

The result of this process was a consistent partition, despite the stochastic nature of the algorithm, in which fewer than 1% of users were partitioned differently in subsequent runs of the full consensus process.

The implementation of the Leiden Algorithm used was the leidenalg Python package created by the original developers of the algorithm at Leiden[31]. Both modularity-based partition and a search over resolutions parameters for the Constant Potts Model were performed with this package. Otherwise, default parameters were used. The Infomap algorithm used the infomap Python package with default settings.

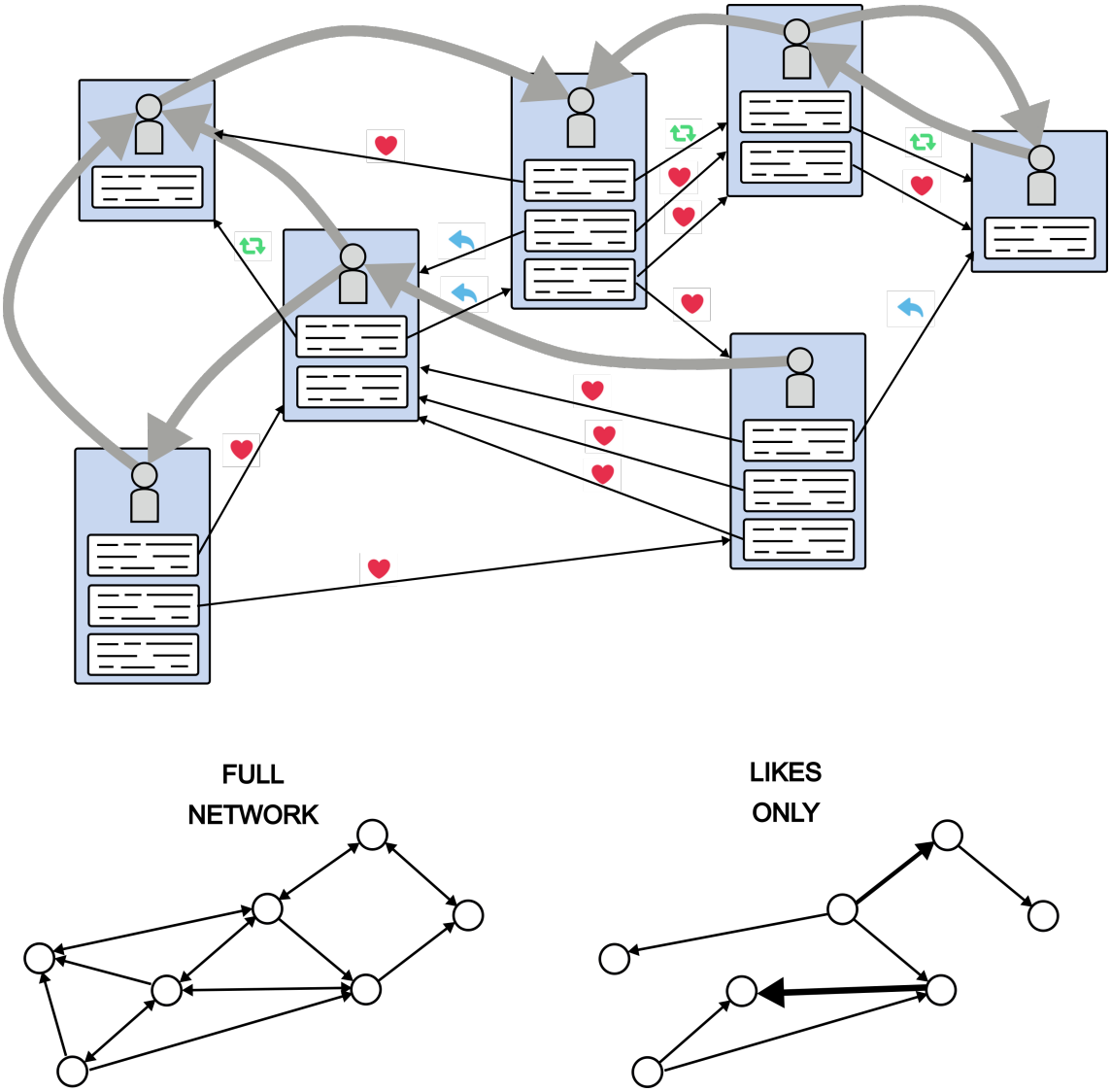

Figure 3.4: Example construction of the user network graph based on tweet connections. Users like, retweet, or follow individual tweets from other users. All arrows follow the influence a user has; the gray arrows represent users following other users and point towards the user who is following. Arrows which point from a user's tweet, shown in white boxes, point towards the user who liked, replied, or retweeted. Arrow thickness in the final graphs below indicate the weight of the connection.

3.4 Classification of Pro- and Anti-NATO Users

To validate the claims that user connectivity is related to user opinion, it was desirable to have a way to estimate the opinion of each of the communities detected regarding the question of whether Sweden should join NATO. The terms ”pro-NATO” and ”anti-NATO” used here are shorthand for those positions. An individual who is in principle favorable to NATO, but believes that it would not be a good idea for Sweden to join would be called ”anti-NATO” in this assessment. Users are given a ”NATO favor score”, S_N indicating their engagement with pro- or anti-NATO tweets.

3.4.1 Assumptions and Caveats

This assessment was done on the tweet level using some simplifying assumptions.

First, users who are clearly in favor of joining NATO or are clearly against are unlikely to have their opinion changed by any event other than the Russian invasion of Ukraine. For this reason, only tweets after the Russian invasion of Ukraine are considered for determining a pro- or anti-NATO score.

Second, users who like a pro- or anti-NATO tweet do so because they share that opinion. Quotes, replies, and retweets do not necessarily show agreement with the original poster, so they do not indicate position.

Third, tweets which contained a hashtag with a clear relationship to the question were interpreted as a pro- or anti-NATO position. The pro-NATO hashtags considered were *#JaTillNATO, #WeAreNATO, #NATONu, #YesToNATO*. The anti-NATO hashtags were *#NejTillNATO, #NoToNATO, #Imperialism, #NATOTroll*. No users employed both pro-NATO and anti-NATO hashtags.

3.4.2 Annotation

Aside from the clear indications, such as hashtags, a subset of the tweets was hand-annotated by different annotators. A suitable number of defined pro- or anti-NATO tweets was needed in order to setup graph connections to give a position estimate to a useful fraction of users. Some degree of manual annotation was needed because even the most obvious combinations of words proved misleading when the entire tweet was not taken into account. The phrase ”vi måste gå med i NATO Nu!” is unambiguous, but one must account for irony:

> Vi måste gå med i NATO nu! Annars blir det inget krig!

Another example of an ambiguous case where the motivation of the reactions needed to be taken into account is this use of the very clear phrase ”jag är emot Nato”.

> Jag är emot Nato men har efter annektering av Krim börjat svänga.

Here, the user clearly states that they are against NATO personally, but their admission that they are starting to change was likely to be ”liked” by users who approved of the change to a pro-NATO position, so an anti-NATO assessment here would not be appropriate.

All annotators were native Swedish speakers with university education in their 20s or 30s. Two were female and two were male. The annotators were given the instructions to rate each tweet with a $\{-1, 0, +1\}$ score indicating if the tweet gave evidence of an opinion against Sweden joining NATO in the user who posted it, no indication of opinion, or evidence of an opinion in favor of Sweden joining NATO in the user who posted it.

Annotators were provided with a randomized sample of the most-liked tweets in the dataset which contained the term ”NATO” by itself or as part of another word. They were provided the text of the tweet, as well as the text of any tweet that the tweet to be scored was made in reply to, to establish context. For example, a tweet that says ”nej, absolut inte” is impossible to score. But if the parent tweet is included, and it reads, ”Kanske dags att gå med i Nato?”, then the tweet to be scored illustrates an anti-NATO position and receives a -1.

Annotators were instructed to give all tweets that were simply reiterating another person’s position a 0, unless they contained words that clearly indicated the users’s opinion of that other person’s stance, for example by saying that ”unfortunately the Social Democrats oppose NATO”, the user is not only reiterating the stance of the Social Democratic party but also reacting to it, so the tweet should be scored $+1$.

3.4.3 Scoring

Approximately 2.5% of the original tweets in the dataset was annotated in the manner specific above. For each user, a ”NATO Score” was determined based on posting and liking patterns.

$$S_N = \sum_i T_i + \frac{1}{2} \sum_j L_j \tag{3.1}$$

Here, T_i is the score given to the i-th tweet posted by the user, and L_j is the score given to the j-th tweet liked by the user. If the final rating was 1 or above, the user was given a pro-NATO rating. if the rating was -1 or below, the user was anti-NATO.

	Tweets Annotated	Users Posting	Users Liking	Total Rating
Pro-NATO	5955	1301	23630	20967
Anti-NATO	3018	1201	12753	8662
Neutral	13834	–	–	–

Table 3.6: User position based on manual annotation of 2.5% of the data.

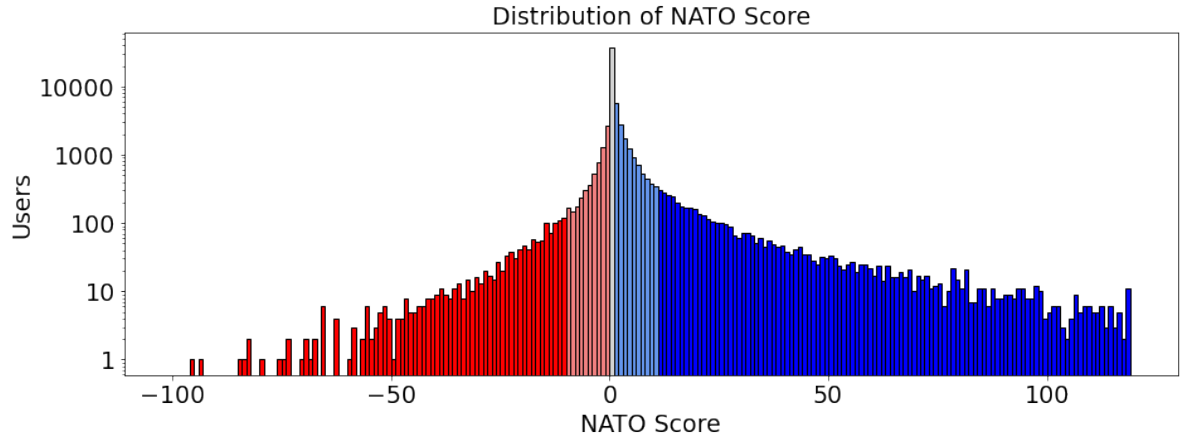

Figure 3.5: Log-scale plot of the distribution of NATO score amongst the users. Scores greater than 10 in either direction are rated ”Strongly” pro- or anti-NATO, and those below the threshold ”Weakly”. The majority of users were ”neutral” in their leaning.

As Figure 3.5 shows, the distribution of user ratings was unimodal–too little of the dataset was annotated to capture a true impression of user opinion, and most users whose engagement was

captured had low engagement overall. Thus, no reliable conclusions about individual users can be drawn from this, but the user distribution gives some indication of a pro- or anti-NATO leaning to a collective. This values give a 2.4:1 ratio of pro-NATO users to anti-NATO users, not too far off from the 2:1 ratio indicated in the previously cited polls. Although here, a minority are in favor and most are neutral, perhaps a consequence of not annotating enough data.

3.5 Text Preprocessing

An important step in the work was to prepare the data so that it was in a form that the text analysis models could use as an input. This pre-processing consists of a number of steps with the end goal of converting the raw text data to a sequence of lexical tokens, each a discrete string of characters with an identifiable meaning.

As previously discussed, the Word2Vec models were used to learn a vectorized word embedding for the words that appeared in the dataset. The lexical tokens that resulted from the pre-processing form the set of ”words” that the models would find an embedding for. It is worth noting here that the tokens do not represent only words–an emoji, for example, is not a word in the traditional sense but carries an identifiable meaning and so it should be classed as a token for processing.

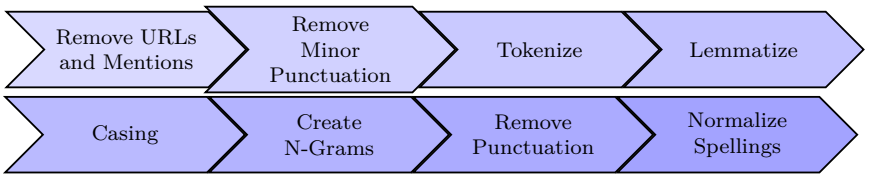

Figure 3.6: The pre-processing pipeline for the Twitter dataset.

Figure 3.6 shows the pipeline used to pre-process the text for use in the Word2Vec model.

Following this preprocessing method, from 859220 tweets 19.9 million tokens were extracted. This set represents the entire Twitter dataset collected here.

3.5.1 Removing URLs and Mentions

Given the nature of Twitter communications, the text of a tweet often contains elements which are unrelated to the linguistic meaning of the tweet. These elements are primarily ”mentions” and URLs. For example:

> @SomeGuyInSweden If you want to remove all URLs and mentions from a tweet, you can look at how Twitter embeds the information https://t.co/JUbfkx9w61

The reference to ”@SomeGuyInSweden” is not relevant to the linguistic meaning of the sentence, so it should be removed from the text. These references will play a role in the construction of user connection graphs, but not for analyzing the text itself. The URLs embedded in the text were removed as well since those also contain no linguistic information. Both these mentions and the URLs are contained in the ”entities” field for each tweet, so the removal process is made simple by removing the mentions and URLs that are listed as ”entities” from the text.

However, it is worth noting that removing mentions can have an effect on some sentence meanings. For example, if a user tweets something to the effect of ”We should support @SomeGuyInSweden in his project,” then the mention takes the place of a noun and has linguistic meaning in the

sentence. This is even more relevant if the mention is of a public figure or organization. However, because the vast majority of mentions were not used in this manner and it would be difficult to isolate the ones that were, it was decided to remove them to clean up the data and reduce the number of tokens.

3.5.2 Removing Minor Punctuation

While punctuation that is important to sentence structure is kept to improve the accuracy of the sentence tokenizer, some minor punctuation that is used in Swedish is removed to avoid erroneously splitting abbreviations and contractions into multiple tokens.

In Swedish, it is common to use a colon (:) in an abbreviation. A reference to a saint, such as Saint Sigfrid of Sweden (Sankt Sigfrid), would often shorten ”Sankt” into ”S:t”. Similarly, references to political parties will typically use a construction such as ”SD:are” for those who vote for, or belong to, the Swedish Democrats party. Twitter users often do not follow conventional grammar, so this is just as often written as ”SDare” or ”sdare”. Removing the ”:” here improves the tokenization accuracy for these structures.

The apostrophe (’) which is used in English to form contractions or the possessive case of nouns is seldom used in Swedish, however it does appear in words borrowed from English as well as names. These were also removed in this step.

Finally, inconsistent hyphenation (-) of words appears in Twitter data. One user may write ”natoland” while another ”natoland” and still another may write ”nato land”. Removing the hyphens in this step allows all three formations of the same token to be treated the same.

3.5.3 Tokenization

Tokenization is the process of breaking down a text string into the previously mentioned lexical tokens. A sentence such as ”Hon har en hund med sig.” should be converted into the series ”Hon”, ”har”, ”en”, ”hund”, ”med”, ”sig”, ”.”, where each word has an identifiable linguistic meaning. Note that often the punctuation is treated as a ”word” in this breakdown.

Various tokenization packages are available, each with strengths and weaknesses. Because this project looks at data from twitter, a purpose-built Tweet Tokenizer in the NLTK package was used to tokenize the tweets[40]. The advantage of this tokenization package is that hashtags and emojis are correctly tokenized without being split. Using the standard NLTK tokenizer or the Stanza default Tokenizer with a tweet containing the following fails to correctly tokenize the second emoji, as there are no spaces between it and the surrounding words.

😂 det var ju nydemokrati🥲deras fel sd fick uppmärksamhet...

The Stanza and NLTK tokenizers return a tokenization consisting of ”😂”, ”det”, ”var”, ”ju”, ”nydemokrati🥲deras”, ”fel”, ”sd”, ”fick”, ”uppmärksamhet”, ”...”. The NLTK Tweet Tokenizer handles this situation well, giving instead a tokenization that contains three separate tokens for ””nydemokrati”, ”🥲”, ”deras”.

Hashtags are also handled more accurately by the NLTK Tweet Tokenizer. Hashtags such as ”#NATO” would often be tokenized into ”#” and ”NATO” by the standard NLTK word tokenizer or the Stanza tokenizer, however the NLTK Tweet Tokenizer correctly identifies ”#NATO” as a single token.

The tweets are tokenized on two levels. First, they are split into sentences, and then each sentence is split into individual tokens. Splitting the tweets into sentences improves the accuracy of the

lemmatizer.

3.5.4 Lemmatize

There are many different forms of the same word that can appear in a text set. In Swedish, verbs inflect with tense (e.g. *sprang* for ”ran”, past tense, or *springer* for ”running”, present tense). Nouns in Swedish inflect with number, case, and whether they are definite or indefinite (e.g. *katt* for a single cat in the indefinite, *katter* for multiple cats in the indefinite, *katterna* for multiple cats in the definite).

Lemmatization is the reduction of the various forms of a word into its lemma, the canonical form of the word. The lemma for *sprang* and *springer* is ”springa”, and the lemma for *katt* and *katterna* is ”katt”.

For word embedding, it is only the lemma of the word that should be considered in the ideal case in this analysis. Correctly connecting all the different forms of a word together gives a clearer picture of the meaning and usage of that term than would emerge if every form were considered a unique and unrelated token.

Several different lemmatization models were tried in this work. The best results were received when lemmatization was done by the Stanza package [41] using a model pre-trained on the Talbanken corpus. The neural network model employed by Stanza attempts to categorize each term by its part of speech and then finds the corresponding lemma. Stanza was found to be significantly slower, but somewhat more accurate, than the other lemmatization tools such as UDpipe.

3.5.5 Casing

The text tokens should all be lowercase. There is no difference in meaning between ”NATO”, ”Nato”, and ”nato”, and there is little consistency in capitalization in tweets. Words that begin a sentence tend to be capitalized, but ”jag” in both ”Jag ska ...” and ”Det var inte jag...” represent the same token regardless of that capitalization.

Performing this before the lemmatization was found to decrease the accuracy of the lemmatizer, as the Stanza Lemmatizer used the capitalization to help categorize a word as a proper noun. For example, depending on the context, the word ”magdalena” might be identified as a noun and lemmatized to ”magdal,” while ”Magdalena” would be recognized as a proper noun and would not be lemmatized.

3.5.6 Create N-Grams

Some words and phrases occur so commonly together that they represent a singular linguistic meaning despite being composed of two words. In a language such as Swedish, where many words are compound words constructed of two other words, this occurs less than in English but some fixes are still necessary. The terms *nato* and *medlem* next to each other should be evaluated the same as the single token *natomedlem*, likewise for any other NATO-related composite terms.

Names that are relevant to the topic, such as Magdalena Andersson or Morgan Johansson, should be considered a single token as well, since there is no linguistic value in separating them. Other 2-grams that are combined into a single token include *artikel* and 5, referring to the mutual defense clause in the NATO agreement, *2* and *%* for the NATO-recommended defense spending target (two percent of GDP), *kalla* and *krig*, referencing the Cold War between NATO and the Warsaw Pact nations, and emoji versions of the letters ”S” and ”E”, ”F” and ”I”, and ”U” and ”A”, which are references to Sweden, Finland, and Ukraine respectively.

The only relevant 3-grams in the text were constructions of the sort *90-talet*, which is how the 1990s are referred to in Swedish. The tokenizers do not consider letters and numbers to be part of the same token so this required fixing after the fact.

3.5.7 Remove Punctuation

After all other processing, punctuation is stripped away. Periods, commas, quotes, and parentheses, etc., are removed from the list of tokens.

3.5.8 Normalize Spellings

Certain terms of interest appeared with multiple spellings. The nature of Word2Vec means that two words, even if they share very similar spelling, will be treated as completely independent tokens. Training a dataset which contains references to ”grey skies” and ”gray cats” may not results in any connection being found between ”grey” and ”gray” as being a color/shade, nor between ”skies” and ”cats”, despite both being things which commonly are gray.

While in some cases it made sense to normalize multiple spellings of a single term, such as a name, or to fix common misspellings in order to connect the multiple uses of the word, in other cases the differences in spelling were important distinctions that motivates keeping them as different tokens.

The Swedish word *medlemskap* was frequently misspelled as ”medlemsskap.” The name of the President of Ukraine was often romanized as ”Zelenskyy”, ”Zelenskij”, and ”Zelenskiy”. The names of Amineh Kakabaveh and Crimea also appeared often with unusual spellings relative to the canonical Swedish form. Swedish-speaking users were also inconsistent with the application of the diacritic in the name ”Erdoğan”, so all references were normalized to ”erdogan”.

Some common nicknames were also standardized in this step. References to the US President as ”Joe Biden” were normalized to ”Joseph Biden”, as the difference was not seen to have any difference to Swedish speakers who have a different culture towards nicknames compared to Americans. The objective was to have as few tokens as possible refer to the same individual when there was no information contained in the differences. Last names were not normalized to full names, as there was not 100% certainty in this kind of matching.

The cases where spellings were not normalized include those in which the novel spelling is believed to contain actual meaning. Writing ”Kyiv” instead of ”Kiev” for the capital of Ukraine is not merely a disagreement in romanization of the original Cyrillic spelling, but rather can be interpreted as a conscious decision to use a romanization of the Ukrainian pronunciation instead of the Russian one, implying a user’s political position. Likewise, the eastern region of Ukraine can be written as both ”Donbass” and ”Donbas”. Even though the difference is most likely a result of the Swedish-speaking users’ unfamiliarity with a foreign term, the differences can reflect a Russian vs. Ukrainian spelling and that distinction could be interesting in the data.

3.5.9 Other Common Processing Steps

There is no universal standard for text pre-processing. It is often task-dependent. Several other operations were considered, and then rejected.

3.5.9.1 Remove Stop Words

Many words in Swedish are so common that they would be expected to appear in nearly every text. Articles such as *ett* and *en*, roughly ”an/a”, and frequently used terms such as *att*, meaning

”that”, quickly become the most common tokens in the dataset and can often add noise to the word embeddings and increase processing time. These terms are called ”stop words”.

It is not always desirable to remove stop words. If all stopwords were removed from the text ”Han tycker att filmen var inte bra”, the result would be ”tycker filmen bra”–a negation in meaningful from ”He did not think the film was good” to ”think film good”.

Furthermore, removing stop words was found to weaken the relationship between certain types of words. Word vectors representing places, for example, were found to be less likely to show the expected relationships when the stop words which could establish a spatial context (till, från, inom) were removed.

3.5.9.2 Remove Singletons

Many of the tokens appear in the text only once. These are called ”singletons”. Word2Vec will not be able to detect any meaningful pattern of use for a single instance of a word, so the word embedding will not be a reliable indicator of the word’s usage or meaning, so their inclusion is not helpful.

However, many of the packages for implementing Word2Vec impose their own minimum frequency when building the word vocabulary, so removing these from the input text was found to be an unnecessary step.

3.6 Pre-Trained Models

The amount of Twitter data collected is not enough to train a Word2Vec model without some pre-training. Each model that was trained in this task represented only one user community or one time step and include roughly one quarter of the data at most. In a similar study of language processing models for Swedish, it was found that datasets of 5 - 15 million tokens were insufficient to train a Word2Vec model from scratch. In that study, the researchers used a model that was pre-trained on a larger selection of similar data and then performed fine-tuning using the dataset of interest, and this project follows the same recommendation [21].

No publicly available pre-trained models were used. Rather, the model was trained on public corpora. While some pre-trained vectors did exist, these could not be fine-tuned without additional restructuring. Publicly available vectors also did not allow different model parameters or pre-processing methods of the text to be explored since the vectors had been trained with a predetermined set of hyperparameters and text pre-processing steps which, and if subsequent fine-tuning used a different set of parameters or pre-processing then optimal results could not be expected.

Keeping in line with general recommendations, and through trial and error, the parameters used for pre-training were word vectors of 300 dimensions, window sizes of 5, and minimum word frequency of 10.

3.6.1 Training Corpora

Word2Vec models were pre-trained on the Göteborgs-Posten 2002-2013 (GP) Corpus, Swedish Wikipedia (W) Corpus, SVT 2001-2013 (S) Corpus, Webnyheter 2001-2013 (WN) Corpus, and the Flashback Övrigt (F) Corpus from Språkbanken[42]. Each token in the corpus was expressed in both the raw and lemmatized form.

To avoid biasing the data with post-2013 information related to Ukraine, the Wikipedia Corpus was scrubbed of all references to the Euromaidan uprising in Ukraine and the subsequent conflict

Corpus	Total Tokens	Vocabulary (Raw)	Vocabulary (Lemmatized)
GP	197.4 M	2.97 M	2.54 M
W	251.9 M	4.52 M	4.17 M
WN	171.0 M	2.53 M	2.16 M
S	63.7 M	1.08 M	0.89 M
F	91.3 M	1.32 M	1.13 M
All	775.3 M	9.15 M	8.40 M
Min 10	758.2 M	1.04 M	0.88 M

Table 3.7: Sizes of the pre-training corpora.

between Russia and Ukraine. All entries from the Flashback corpus from after 2013 were removed. The corpora for the pre-training of the model reflects Swedish language use from 2001-2013.

Multi-word references for names, places, and events in the corpora were collected as a single token. For example, an entry about Swedish Prime Minster Magdalena Andersson would have the name compressed to a single token, ”magdalenaandersson” for training. This allowed these 2-grams to be treated as a single token, which would correspond to the 2-grams created by the Twitter data pre-processing methods.

Table 3.7 gives a breakdown for the tokens and vocabulary sets for the corpora. The total number of tokens in the pre-training corpus is 759 million. Considering only tokens which appear a minimum of 10 times gives about 1.05 million unique tokens in the raw text, and 0.89 million in the lemmatized text. The huge number of tokens which appear fewer than ten times consist of numeric entries (such as finishing times in a race in the GP corpus or heights of mountains in the W corpus), infrequently mentioned names of people or places,

The primary reason for using such a large corpus was to improve the quality of the pre-trained model. With sufficient input data, the particulars of the data cleaning and model parameters were less important. It was desirable to have a neutral word embedding for many terms which would only become commonly spoken about in Swedish media during the Russian invasion. In general, for tokens which appear at least 10 times in both the pretraining corpora and the Twitter dataset, 93% of vocabulary that appears in the raw Twitter dataset appears the pretraining corpora, and 80% of the vocabulary that appears in the lemmatized dataset does. The remaining tokens from the Twitter dataset which do not appear in the pretraining corpora are emojis, spelling errors, foreign words captured from Tweets which contained multiple languages (or were improperly flagged as Swedish by Twitter), very specific compound words, and numbers. However, around 95% of all tokens in the Twitter dataset belong to the shared vocabulary.

Terms that are relevant to the NATO debate in Sweden appear much more frequently in the Twitter dataset than in the general corpora because these terms formed the basis of the search that the Twitter dataset was built from. Table 3.8 shows the frequency of appearance of a few terms of interest for the NATO discussion. Those terms with the greatest frequency are expected to be pulled out of the original training context more quickly than less frequent terms, and those terms which appear more often in the Twitter dataset than in the original dataset are expected to have their word embedding reflect their context in the Twitter dataset more strongly. The CBOW architecture is expected to be better suited for the high-frequency terms, whereas the low-frequency terms are expected to have a more meaningful embedding from the Skip-gram model. What becomes difficult here is for either model to learn a useful embedding for the term *nato*, as its frequency is as high as the frequency of common Swedish articles *att* and *en*.

Token	Pre-Training Frequency	Twitter Frequency		Token	Pre-Training Frequency	Twitter Frequency
nato	1.01×10^{-5}	1.12×10^{-2}		sverige	8.12×10^{-4}	5.80×10^{-3}
ryssland	9.89×10^{-5}	3.73×10^{-3}		ukraina	2.39×10^{-5}	2.96×10^{-3}
putin	8.16×10^{-6}	2.24×10^{-3}		usa	3.52×10^{-4}	2.04×10^{-3}
finland	2.20×10^{-4}	1.83×10^{-3}		krig	5.06×10^{-5}	1.50×10^{-3}
natoansökan	2.76×10^{-9}	4.00×10^{-4}		neutralitet	1.18×10^{-6}	2.55×10^{-4}
suveränitet	1.93×10^{-6}	1.55×10^{-4}		alliansfritt	1.94×10^{-7}	7.71×10^{-5}
natomedlem	9.75×10^{-8}	7.40×10^{-5}		azov	7.54×10^{-8}	4.33×10^{-5}
kyiv	1.02×10^{-8}	1.41×10^{-5}		säkerhet	3.57×10^{-5}	4.82×10^{-4}
att	1.48×10^{-2}	3.13×10^{-2}		en	1.61×10^{-2}	1.29×10^{-2}

Table 3.8: Comparison of frequency of NATO-related terms in the Twitter dataset and n the pre-training corpora.

3.6.2 Validation

Several tests were performed on the pre-trained word vectors to verify that the model had realistically captured the use of the Swedish language. For each of the measures, the source paper for the metric is used as a benchmark.

The first measure was the QVEC-CCA score[43], which measured the relationship between words of the same course semantic categorization, based on the SALDO lexicon[44]. Previous measures of this using Swedish text achieved scores of $0.3496 - 0.3516$ for CBOW models with a window size of 5-10, $0.3517 - 0.3555$ for skip-gram models [23].

The second measure was a measure of word similarity based on the SuperSim test[45]. Hand-annotated pairs of words are scored for this similarity, and the ranking is compared between the annotations and the cosine similarities of the resulting word vectors using Spearman's ρ according to:

$$\rho = \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{3.2}$$

Where d_i is the difference in rank position between the average score of the annotators and the cosine similarity of the word vectors for a word pair. The original paper returns scores of up to 0.496 for Word2Vec models trained on a billion-word Swedish corpus.

Also run was an analogy. Sets of four words, $wv_1 : wv_2, wv_3 : wv_4$, are given in such a way that the vector math of the first three words of the form $wv_1 - wv_2 + wv_4$ should result in the third word. For example, given the analogy "kung:man,drottning:kvinna," the vector summation *kung - man + kvinna* is expected to result in *drottning*. The original paper [46] gave accuracy of up to 60.4% for skip-gram and 54.4% for CBOW Word2Vec models (both using negative sampling). The table gives both the average value (out of 1.0 for perfect agreement of all terms), as well as the accuracy (the number of pairs for which the correct answer had the highest cosine similarity, regardless of the absolute value).

The final test was to check how the model performed on estimating word forms (e.g. *står:stod,sover:sov*), or analogies based on the inflection of the word due to tense or number. This score presents the ratio of tests in which the trained model returned the correct result given the analogies based on word form. The lemmatized models could not perform this task.

Table 3.9 shows the resulting scores for models trained on the pre-training data. The results are comparable to the benchmarks, indicating a good degree of appropriate training. The validity of the

Model	Vocabulary Size	QVEC-CCA Score	SuperSim Spearman ρ	Analogy Accuracy	Word Form Accuracy
Word2Vec SGNS (Raw Text)	1.05 M	0.375	0.605	0.570	0.599
Word2Vec CBOW (Raw Text)	1.05 M	0.374	0.546	0.425	0.674
Word2Vec SGNS (Lemmatized)	898 K	.0381	0.602	0.608	N/A
Word2Vec CBOW (Lemmatized)	898 K	0.380	0.595	0.462	N/A

Table 3.9: Validation of models trained on the pre-training corpus.

Corpus	Total Tokens	Vocabulary (Raw)	Vocabulary (Lemmatized)	New Tokens (Raw)	New Tokens (Lemmatized)
All	19.7 M	294.5 K	234.0 K	94.8 K	100.7 K
Min 10	19.2 M	42.5 K	31.0 K	3.1 K	5.9 K

Table 3.10: Sizes of the Twitter dataset.

model is important only in providing a base-line vectorization of words that appear in the Twitter dataset; the downstream task does not require perfect agreement to an arbitrary set of vectors.

3.6.3 Vocabulary Comparison

The entire Twitter dataset used had 42462 unique tokens when considering only those which appeared more than ten times, and 30973 unique lemmas. Table 3.10 shows the vocabulary sizes of the Twitter dataset. Of the 19.5 million tokens in the dataset, 98.4% of them are appearances of tokens that belong to the vocabulary of the pre-training data.

Retraining the model with a fine-tuning dataset has the potential to overwrite the word embedding vectors for those vocabulary entries which appear in the new dataset. New vocabulary entries will be added, and because these new entries do not appear in the pre-training dataset, they will be initialized at random and the only context that they will have is associated with the text in the fine-tuning dataset. Other tokens that do not appear in the Twitter dataset may lose their association with the new words.

The setup of this training process is highly dependent on the downstream task. In some tasks this is problematic, because if a token such as ”världsorganisation” is found to be one of the more closely associated tokens with ”nato” in the original corpus, but this token does not appear at all in the Twitter dataset, then retraining on the twitter dataset alone has the potential to to pull the common vocabulary terms out of their original context. In this case, however, pulling the text out of its original context is not undesirable. As in the previous study of Swedish parliamentary motions[21], we are interested most in the differences in word embeddings between user groups–if text is not used by any of our user groups, it is irrelevant anyway, so the separation from previous embeddings is not a loss.

3.7 Training the Models on the Twitter Dataset

The Twitter data set was broken down into several corpora, each representing either a time step in the development of Sweden's NATO application, a user community or sub-community, or an estimate pro- or anti-NATO leaning. A unique model was fine-tuned using the text from each corpus, based on the pre-trained base model with the corresponding setup (e.g. a pre-trained CBOW model based on lemmatized text was used for lemmatized tweets).

The Word2Vec models were created using the Gensim Python package[47]. Each corpus was fine-tuned on the CBOW and skip-gram architectures with and without lemmatizing for 160 epochs with a window size of 10, a vector size of 300, minimum frequency of 10, initial alpha of 0.01 and all other hyperparameters default. The loss after 160 epochs was stable enough in the larger communities that no further training appeared to be necessary.

Each model architecture was run 5 times on the corpus from each community or timestep and the top ten most similar words by average cosine similarities were assembled for each of the words in a list of words of interest. Word embedding vectors were also stored for each of the words of interest for each model architecture, community, and time step.

Only cosine similarities within communities was considered; a raw comparison of vectors is unlikely to yield useful results, as the entire word embedding space can have arbitrarily rotated during training such that the vectors for a given token no longer point in the same direction, but the cosine similarity between that token and other tokens remains the same across both communities.

3.7.1 Keywords

The results of the model training is a set of vector word embeddings, taken from the W matrix in the Word2Vec model. With a vocabulary of hundreds of thousands of words, an exhaustive comparison of all terms to find those with meaningful semantic shift is not possible. Instead, a subset of the vocabulary of 8000 keywords deemed relevant was defined as a starting point. These are words whose comparison to other words were be tabulated.

3.7.1.1 Known Terms

The most common words were considered, as they are most relevant to the topic of conversation. References to *nato* or *sverige* are obvious choices for keywords.

The list also contains words which, based on prior knowledge of the topic, are believed to be most likely to return polarizing results. Terms related to political parties and the main political figures in the discussion, such as *putin*, *kakabaveh*, and controversial areas of discussion such as *minskavtal*, *statskupp*, or *alliansfrihet*, and all variations thereof, are also added.

3.7.1.2 Unknown Terms

A key part of this research is the development of automated methods to find terms which are meaningful to a discussion but whose importance is unknown to researchers. One approach to this is to compare the distribution of word frequencies within communities. The previously stated hypotheses assume a contentious topic with at least two polarized sides which are expected to use language differently. Without knowing which communities will belong to either side, one of the most straightforward ways to find words with polarized usages is to look for multimodality in the frequency distribution of all words across communities. If a number of communities represent one side of an argument and a number of communities another side, the distribution of any word which

is strongly indicative of a particular side of the issue is likely to be different across the two groups of communities.

A search for multimodality can be done using Hartigans' Dip Test[48]. The Dip Test looks for a characteristic dip in the cumulative probability distribution function for a variation, as in Figure 3.7. The depth of the dip can be used to characterize the probability of multimodality in a function and can assist a search for keywords.

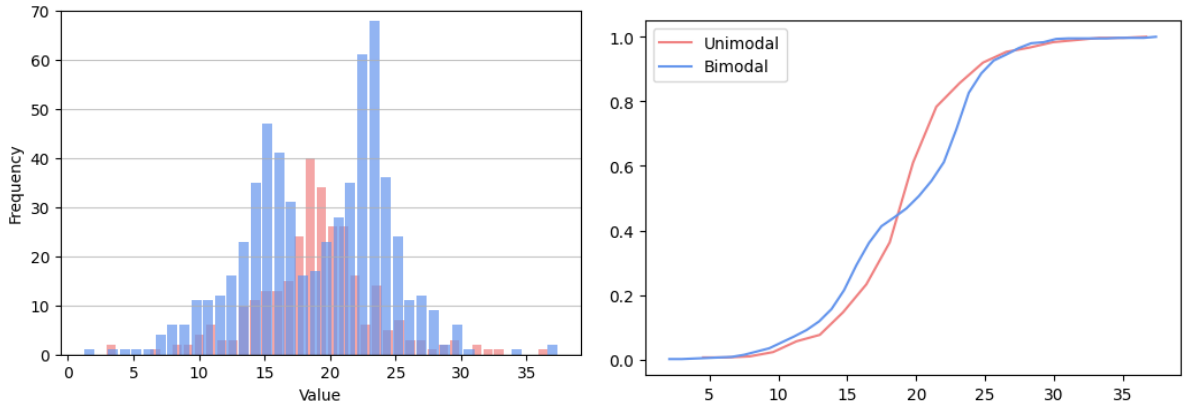

Figure 3.7: A "dip" in the cumulative probability distribution function for a bimodal distribution compared with a unimodal distribution's cumulative probability curve.

The distribution of words in each of the corpora assembled for the communities detected by the community detection algorithms was submitted to the Dip Test. Words which appeared frequently in some communities but seldom in others could signify coded words, shibboleths, or novel uses of words. The words with the greatest degree of multimodality were added to the list.

The final list was assembled by finding the most similar 20 words to each word in the original list by running the entire Twitter dataset through the Word2Vec model. The result was a list of approximately 8000 keywords.

3.7.2 Evaluating Cosine Similarity and Words of Interest

The trained models yielded one 8000×300 set of vectors and architecture and 8000 sets of the most similar words for each of the keywords for each community, timestep, or NATO-score classification. An extensive investigation was performed to find word relationships which were indicative of one particular side of the NATO debate, isolated to particular communities, or otherwise of interest. "Words of Interest" is used here to refer to those keywords whose usage presented interesting patterns which could help to distinguish the ideological position of user communities.

Generally, speaking, there were two method used. First was a comparison of the ten most similar words across communities to see if those communities fell into any dichotomy where certain relationships were mutually exclusive. The second was to look at cosine similarities between words to see if a pattern emerged across communities.

At times, the communities were sorted by their average NATO-score to detect any correlations between word cosine similarities and user position.

The source for words and relationships that were explored was:

- The most common words in the dataset
- The terms expected to be most divisive from prior knowledge
- Words for which any related terms appeared frequently within the top 5 most similar words within the communities with the most strongly pro- or anti-NATO scores, but not at all within

the top 10 most similar words in the opposing communities
- Words for which any related terms appeared as one of the top 5 most similar words among either the pro- or anti-NATO user set as a whole, and not within the top 10 most similar of the other
- Cosine similarities which showed the greatest multi-modality according to the Dip Test
- Cosine similarities which in some communities had increase substantially from the base model while decreasing substantially in other communities
- Cosine similarities which held a strong outlier, where one community had a relationship above a threshold while all others were below a lower limit

Once terms and relationships were defined, they were manually investigated.

4

Results

This chapter covers the results of the community detection, both in terms of language use and the NATO opinion score derived from the manual annotation, and presents a detailed analysis of the language use patterns of the users in the dataset based upon the most similar words by cosine similarity for those words selected by methods discussed in the previous chapter.

4.1 Community Detection

Between the 8 different partitions of the user base, 76 communities of users with more than 300 members were found. The consensus clustering approach led to communities that were stable across multiple runs of the various community detection algorithms.

4.1.1 Community Structure

Tables 4.2 - 4.1 show the sizes of the major communities and notable subcommunities detected using the various community detection algorithms. The Leiden Algorithm yielded a set of five major communities and several minor communities in the full network, and four major communities with several minor communities in the likes only network. The Infomap algorithm returned two major communities in the full network, and one major community in the likes only network. Both returned several minor communities with a few thousand members. The size of the communities detected by the Constant Potts Model was a function of the resolution parameter. The tendency of the Constant Potts Model was to detect numerous smaller communities of a couple hundred users, representing a few thousand tweets at most, which would collapse into a single large community of 50 000 or more users past a certain resolution threshold. A satisfying trade-off between many very small communities and one extremely large one was seen with a resolution parameter of 0.05, so that value was used for further analysis.

Full Network

Community	Users	Tweets	Tokens
0	4343	296941	6.71 M
1	3588	138487	3.28 M
2	1036	3328	76 K
3	931	10023	241 K
4	723	36966	863 K
5	611	49	N/A
6	407	4416	N/A

Likes Only

Community	Users	Tweets	Tokens
0	4193	169438	6.39 M
1	1768	170493	2.62 M

Table 4.1: The sizes of each of the user communities and the resulting text corpora from the Constant Potts Model partition.

Full Network

Community	Users	Tweets	Tokens
0	15691	382129	8.59 M
1	13561	197630	4.68 M
2	3724	105159	2.49 M
3	2970	6788	149 K
4	2506	15839	378 K
5	2424	28950	686 K
6	1017	1970	49 K

Likes Only

Community	Users	Tweets	Tokens
0	11072	169438	4.03 M
1	5970	170493	3.99 M
2	4978	41659	970 K
3	3523	106046	2.50 M
4	2427	35389	1.34 M

Table 4.2: The sizes of each of the user communities and the resulting text corpora from the Infomap partition.

For each community detection method (Leiden, Hierarchical Leiden, Infomap, and Constant Potts), and each of the two network setups, a short assessment was done to attempt to quantify the quality of the communities detected. This methodology followed that in [26] and the results received were comparable. For each partition, the directed modularity was determined as in (2.3).

In addition to modularity, several other measures were taken.

Expansion is the total weight of edges leaving the community, $|E_c^{in}|$ divided by the number of nodes in the community. The lower this value is, the better quality the community is. The maximum value of expansion from all communities is given for each partition.

$$\frac{|E_c^{in}|}{n} \tag{4.1}$$

Contraction is the total weight of edges inside the community divided by the number of nodes in the community. The higher this value is, the better quality the community is. The minimum value of contraction from all communities is given for each partition.

$$\frac{|E_c^{out}|}{n} \tag{4.2}$$

Conductance is the proportion of the weight of edges connecting to nodes outside the community, $|E_c^{out}|$, relative to the total weight of edges that the nodes in the community are members of. The closer this measure is to 0, the more isolated the community is. For each partition, the maximum value of conductance of a community (i.e. worst case) is given.

$$\frac{|E_c^{out}|}{|E_c^{out}| + |E_c^{in}|} \tag{4.3}$$

A summary of these results for each of the partitions is given in Table 4.4. Overall, the Leiden and Infomap Algorithms performed roughly the same on the full network. The Leiden and Hierarchical Leiden algorithms on the likes only network returned weaker communities compared to other partitions. In terms of contraction, expansion, and conductance, Infomap's partition of the likes only network was of similar-or-better quality to the full network partitions. This is likely a result of the fact that the likes only network was not only sparser, but also formed by few nodes with many outgoing edges towards many nodes with only incoming edges. The "flow" modelling of the Infomap partition compared with the modularity, essentially a density comparison, was better suited for structures like this.

The Constant Potts Model returned a partition with a poor modularity and expansion. The communities that were found were small; many of which contained only one or two users. The poor

Full Network

Community	Users	Tweets	Tokens
0	15677	219446	5.20 M
0	4747	36690	870 K
1	4643	60008	1.46 M
2	2764	32241	745 K
3	2504	83764	1.95 M
4	935	6741	83 K
1	14801	243275	5.76 M
0	7541	68852	1.61 M
1	4753	140430	3.34 M
2	2199	30756	720 K
2	11950	318404	7.17 M
0	4668	48229	1.06 M
1	2997	104673	2.46 M
2	2313	110398	2.42 M
3	1923	55097	1.20 M
3	10574	19083	387 K
0	4446	6759	135 K
1	2333	3737	79 K
2	1441	3365	55 K
3	679	913	21 K
4	537	3695	53 K
5	394	266	N/A
4	7438	38383	876 K
0	2298	21145	504 K
1	2190	7275	155 K
2	305	1309	27 K
5	872	901	22 K
6	677	70	N/A
7	379	7224	168 K

Likes Only

Community	Users	Tweets	Tokens
0	17729	224368	5.31 M
0	6467	61345	1.47 M
1	4981	72598	1.78 M
2	2913	68382	1.55 M
3	1675	10630	258 K
4	985	9014	101 K
1	14629	244591	5.76 M
0	3291	9437	209 K
1	2337	28523	661 K
2	2195	32561	812 K
3	1954	44240	1.01 M
4	1645	36607	729 K
5	1470	75428	1.54 M
6	1142	9681	103 K
2	14399	211620	4.60 M
0	5996	37738	830 K
1	4796	57178	1.29 M
2	3538	116649	2.48 M
3	4728	138672	3.34 M
0	1044	16981	404 K
1	1018	12682	289 K
2	990	43102	1.08 M
3	704	29602	686 K
4	654	30137	805 K
4	507	43	N/A
5	346	1835	42 K

Table 4.3: The sizes of each of the user communities, subcommunities, and the resulting text corpora from the Leiden and Hierarchical Leiden partition. Several of the communities had too few tweets to use for any meaningful training. Retweets are not considered here, as they do not contain unique text, so it is possible for some communities to have more users than tweets if the users only retweeted content instead of posting their own.

Partition	Number of Communities	Modularity	Conductance	Expansion	Contraction
Leiden (Full Network)	54	0.382	0.569	223	31.1
Hierarchical Leiden (Full Network)	139	0.268	0.574	198	15.1
Infomap (Full Network)	2 623	0.343	0.608	206	56.6
Constant Potts (Full Network)	20 423	0.203	0.415	591	224
Leiden (Likes Only)	320	0.328	0.89	196	0.14
Hierarchical Leiden (Likes Only)	409	0.194	0.875	167	0.15
Infomap (Likes Only)	4 120	0.197	0.625	104	50.2
Constant Potts (Likes Only)	23 771	0.127	0.629	916	186

Table 4.4: Results for partition validation metrics.

scores from this analysis show that the Constant Potts Model does not find clearly separated communities; the partition is based upon a somewhat arbitrarily chosen parameter. Few user communities detected with this algorithm were large enough to consider for language analysis.

4.1.2 Community Scoring

To determine if there were a relationship between pro-NATO or anti-NATO opinions and language use, an attempt was made to characterize the general opinion of Sweden joining NATO of each community. This utilized the ”NATO Score” from the manual annotation. Figures 4.1 - 4.4 illustrate the communities, their connection to each other, and the distribution of the NATO score of the communities’ members. Especially of note is how the pro- and anti-NATO clusters within the communities found by the Leiden algorithm are broken out into subcommunities when using the Hierarchical Leiden algorithm instead. That these users have connections between them that are recognized and preserved by different partitions suggests that the structure of the network is better represented by these smaller communities than by the larger top level communities, and that the top level partition is assigning these communities to different supercommunities which are too loosely connected to have a real meaning.

It is clear from the figures that the communities are more homogeneous when considering only ”likes” connections. This is most probably a result of how the score is defined; users’ scores are determined by tweets that they post and that they like. If the score were determined based upon following known pro-NATO or anti-NATO users, then we may see a greater concentration of pro- or anti-NATO users in the communities found in the full network, which utilized following information to build connections.

No community has a majority of users with an anti-NATO score on the top level Leiden or in the Infomap partition, even in the likes only network, though Community 3 in the Infomap partition of the likes only network comes closest. Several communities in the top level Leiden have a majority of pro-NATO leanings, as overall users seem to be either indifferent or pro-NATO leaning.

For consideration of the differences in language usage, Table 4.5 gives the strongest ”pro-NATO” communities and Table 4.6 gives the strongest ”anti-NATO” communities. Pro-NATO positions were more pronounced overall; the weakest pro-NATO leaned community was still more decided pro-NATO than most of the strongest of the anti-NATO communities were in the opposite direction, and the strongest anti-NATO community was much smaller than most of the pro-NATO communities. Because these communities are derived from different partitions, users may appear in more than one of these communities.

The Constant Potts Model detected the most clearly pro- and anti-NATO communities. Though most of the communities found by this method were very small, those of substantial size, especially in the likes only network, were more strongly oriented towards one side or the other than the communities found in the other community detection algorithm. Many of the smaller communities, those of 100-200 users, were also more homogeneous in NATO score than communities found in the other algorithms, though they were too small to be used as input for the Word2Vec model.

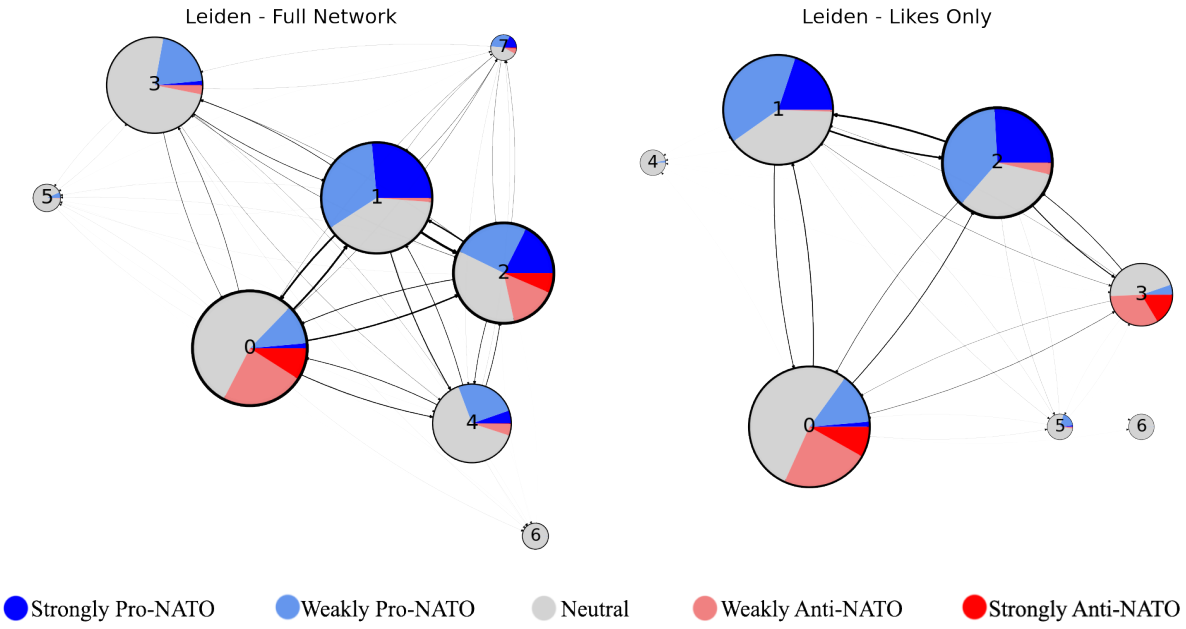

Figure 4.1: The NATO scoring of communities detected using the Leiden Algorithm. The width of the lines indicates the number of connections between communities, and the width of the border of each community represents the number of connections within the community. The size of each node is proportional to the number of users included in the community.

Partition	Community	Users	Pro-NATO Users	Anti-NATO Users	Average Score	Corpus Size
Pro-NATO						
Leiden Full Network	1	14801	59.1%	1.2%	11.4	5.76 M
Leiden Likes Only	1	14629	59.8%	0.5%	8.7	5.31 M
	2	14399	63.8%	3.5%	10.3	4.6 M
Hierarchical Leiden Full Network	1,1	4753	64.9%	0.9%	15.4	3.34 M
	1,2	2199	67.7%	1.2%	16.2	720 K
	2,2	2313	70.7%	6.4%	21.9	2.42 M
Hierarchical Leiden Likes Only	1,1	2337	65.6%	0.2%	9.1	661 K
	2,1	4796	75.4%	0.4%	17.4	1.29 M
	2,2	3538	62.9%	4.6%	12.2	2.48 M
Infomap Full Network	2	3724	61.3%	0.8%	12.45	2.49 M
Infomap Likes Only	1	5970	62.0%	1.0%	12.7	3.99 M
Constant Potts Full Network	4	723	79.4%	0.8%	23.8	863 K
Constant Potts Likes Only	0	4193	92.3%	4.6%	46.7	6.39 M

Table 4.5: The sizes and NATO-scoring of each community that is considered as a ”pro-NATO” community.

Partition	Community	Users	Pro-NATO Users	Anti-NATO Users	Average Score	Corpus Size
Anti-NATO						
Leiden Likes Only	3	4728	5.3%	49.3%	-4.3	3.34 M
Hierarchical Leiden Full Network	0,1	4643	4.7%	52.5%	-4.7	1.46 M
	2,1	2997	4.9%	58.5%	-5.9	2.46 M
Hierarchical Leiden Likes Only	0,1	4981	2.0%	59.0%	-5.5	1.78 M
	3,1	1018	2.6%	66.5%	-5.2	289 K
	3,3	704	5.2%	54.0%	-5.8	686 K
Infomap Likes Only	3	3523	8.3%	55.7%	-4.8	2.50 M
Constant Potts Likes Only	1	1768	13.5%	76.9%	-14.5	2.62 M

Table 4.6: The sizes and NATO-scoring of each community that is considered as an ”anti-NATO” community.

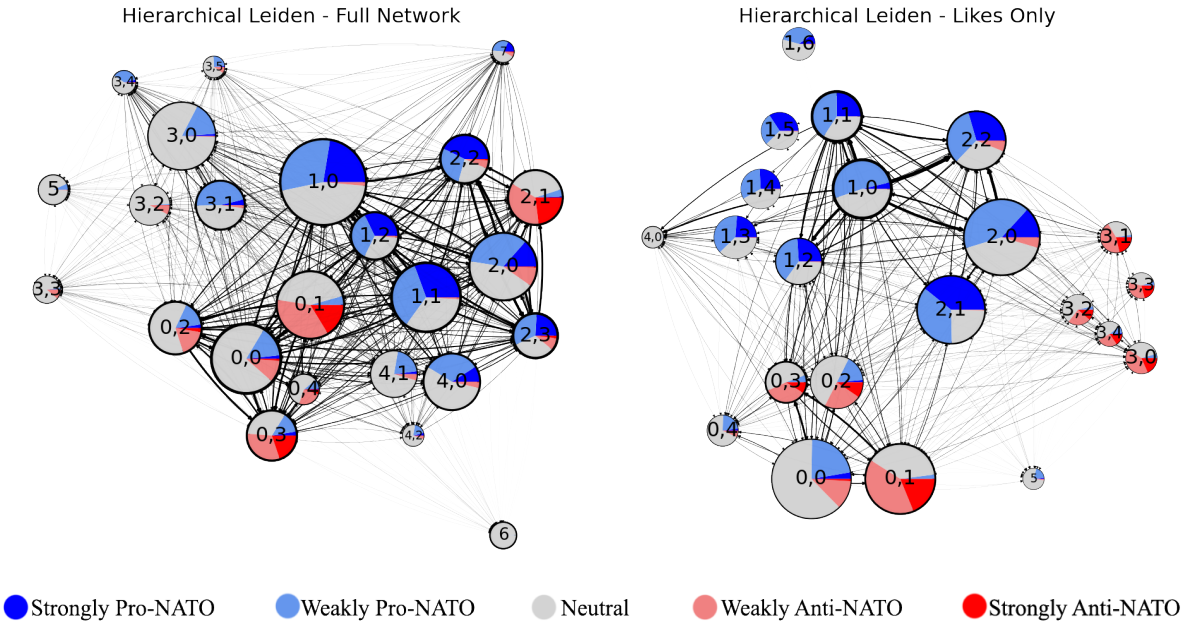

Figure 4.2: The NATO scoring of communities detected using the Hierarchical Leiden Algorithm. The width of the lines indicates the number of connections between communities, and the width of the border of each community represents the number of connections within the community. The size of each node is proportional to the number of users included in the community.

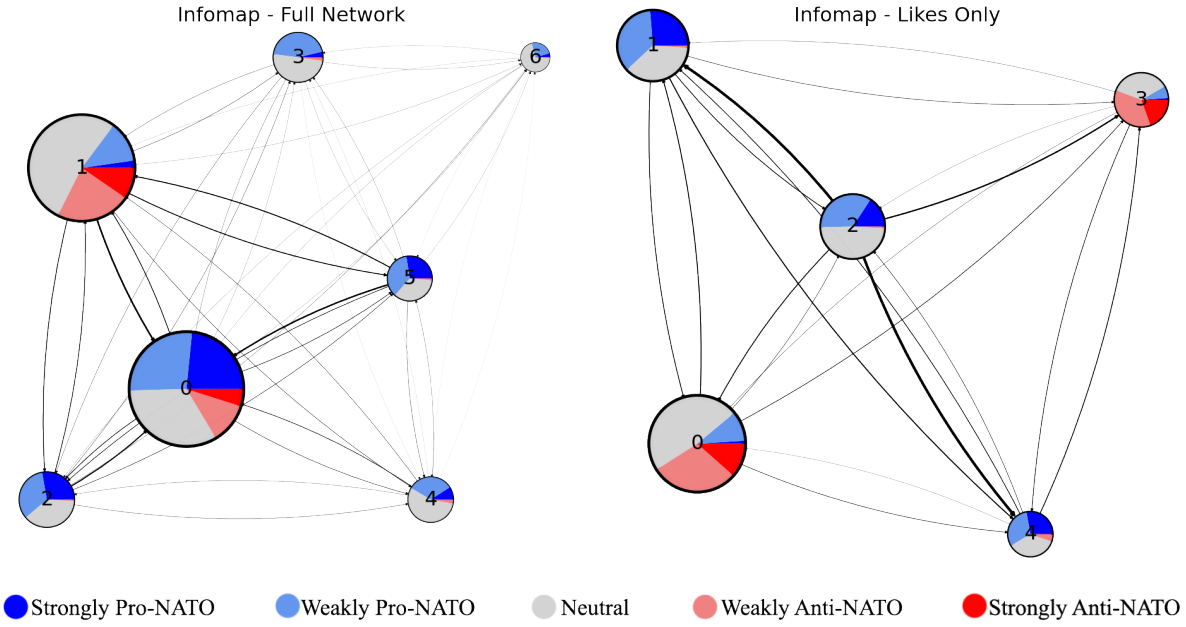

Figure 4.3: The NATO scoring of communities detected using the Infomap Algorithm. The width of the lines indicates the number of connections between communities, and the width of the border of each community represents the number of connections within the community. The size of each node is proportional to the number of users included in the community.

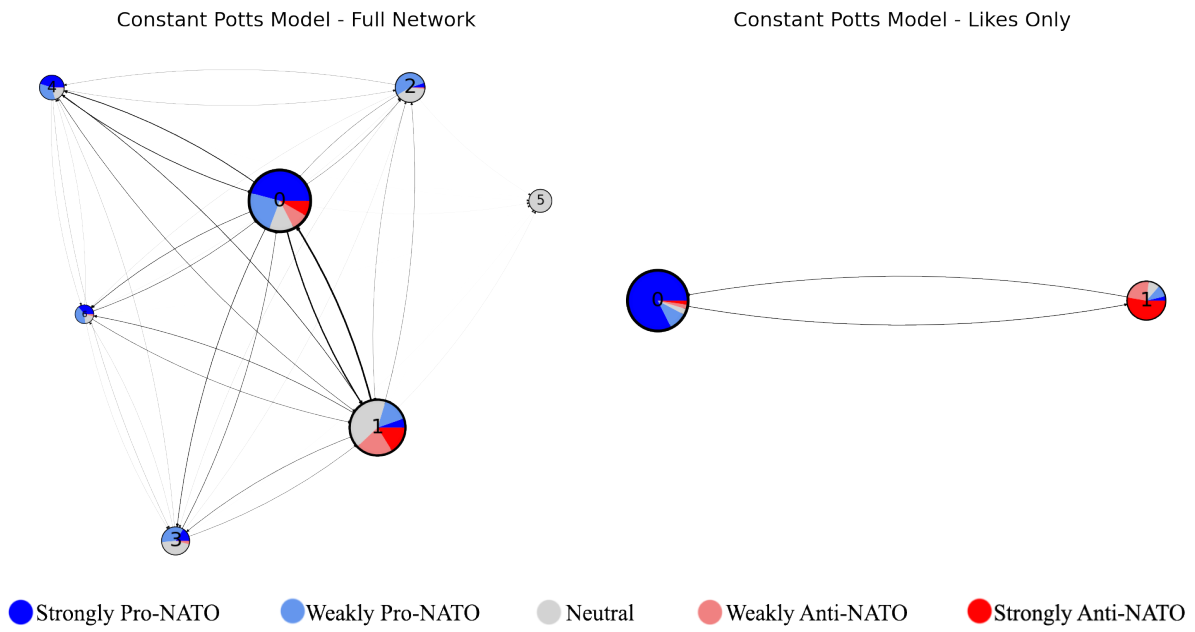

Figure 4.4: The NATO scoring of communities detected using the Constant Potts Model. The width of the lines indicates the number of connections between communities, and the width of the border of each community represents the number of connections within the community. The size of each node is proportional to the number of users included in the community.

4.2 Word Embeddings

4.2.1 General Results

Table 4.7 contains some of the words of interest coupled with the the most similar words for each as determined by a model trained on the given corpus. A general shift towards the context of the Russia-Ukraine conflict is observed in the word embeddings for models trained on the Twitter data when compared to the word embeddings of the pre-trained model. Note that a word which is not among the most similar words of the community can still have a higher cosine similarity than in another community where the word is one of the most similar words if a stronger relationship exists with other terms.

For very common terms, there is little distinction between groups. The terms *nato* and *sverige* appear most commonly in the dataset after stop words such as *att* or *en*. These words are most strongly associated with each other as a result. The term *nato* only appears in the pre-training data 10 575 times, while it appears 236 435 times in the Twitter dataset, leading to a large bias towards the new context. As the users are Swedish, the term *vi* becomes the second most commonly related term to *sverige* after *nato*. Also closely associated are *finland* and *eu*.

Terms related to the context show a broad change from the pre-trained data, but also not much distinction among groups. Most related to the term *krim* in the pre-trained model are *krimhalvön* (omitted in Table 4.7 for being a synonym with no distinction), *kertjhalvön*, the Crimean Peninsula and the smaller Kerch Peninsula on the eastern end of Crimea, and *kiev*, the capital of Ukraine. The Twitter data instead shows a stronger association with *donetsk* and *donbass*, regions that, like Crimea, are contested by the Russian Federation. Also associated with *krim* is *abchazien*, a region in the Caucuses that unilaterally declared independence from Georgia with Russian backing in a similar (and equally unrecognized by the international community) manner as Crimea had done with Ukraine. Terms such as *sydossetien*, *nagornokarabach*, *golanhöjderna*, and *transnistrien* appeared as well, refering to South Ossetia, Nagorno-Karabach, and the Golan Heights, all disputed territory in the Caucauses region or the Middle East, and Transnistria, another post-Soviet frozen conflict zone.

The inclusion of *kosovo* by only one of the groups is of special interest here, as Kosovo is often used as a precedent for legitimisation of the annexation of Crimea by Russia, as The Republic of Serbia does not recognize the indepedence of western-backed Kosovo in a parallel to Ukraine not recognizing the independence of Russia-backed Crimea. The other conflicts associated with *krim* here either have Russia as the backer of the breakaway republic or do not involve Russia at all.

A strong associate between the masculine pronoun *han* and *putin* arises in some of the Twitter user groups. This association stems from the focus on Vladimir Putin's actions in the context of the war in Ukraine. A typical example of this usage is:

> Han kan ju inte anfalla med kärnvapen som svar på sanktioner. Det är ju helt orimligt att Väst / USA / Nato bara sitter och tittar på när stridsvagarna rullar in i ett grannland.

Further association between *ryssland* and *putin* shows that the leader of the Russian Federation is used interchangeably with the Russia Federation by the users. It also shows the arising of a moniker, *putler*, meant to draw association between Russia's invasion of Ukraine and Germany's invasion of European countries during the Second World War. This is an example of a novel word that is discovered by this process.

Likewise, the feminine pronoun *hon* is found to be among the most similar to the word embedding

	nato	sverige	putin	magdalena andersson	krim
Base Model	försvarsalliansen eu	europa norden	medvedev bush	pär nuder anders borg	kertjhalvön kiev
Pre-Invasion	ryssland sverige	nato ryssland	han ryssland	pär nuder anders borg	ukraina dontesk
Post-Invasion	sverige eu	nato vi	han ryssland	statsministern hon	donbass dontesk
Pre-Application	sverige #nato	nato vi	han putler	statsministern fredrik reinfeldt	krimhalvön georgien
Post-Application	sverige turkiet	nato turkiet	erdogan ryssland	hon statsministern	donbass ukraina
Leiden 0	sverige finland	nato vi	ryssland han	statsminister stefan löfven	abchazien donetsk
Leiden 1	sverige #nato	nato vi	han ryssland	statsministern hon	donbass abchazien
Leiden 2	sverige finland	nato vi	han erdogan	statsministern hon	donbass kosovo
Leiden 3	finland sverige	finland nato	saakasjvili erdogan	pär nuder andersborg	balkanhalvön bessarabien
Infomap 0	sverige finland	nato vi	han ryssland	statsministern hon	abchazien donetsk
Infomap 1	sverige findland	nato vi	ryssland han	statsministern stefan löfven	donbas abchazien
Infomap 2	sverige finland	nato ryssland	han ryssland	stefan löfven pär nuder	donbas abchazien
Infomap 3	finland eu	finland tyskland	medvedev saakasjvili	pär nuder anders borg	kiev kertjhalvön

Table 4.7: The words with the top two cosine similarities to the keyword for each of the largest communities using the CBOW model and unlemmatized text from the full network setup. Here, words with the same lemma but different forms (e.g. *alliansfria* for *alliansfritt*) and full names (e.g. vladimir putin for *putin*) are excluded.

for *magdalena andersson*, also giving an indication that the (now) former Swedish Prime Minister is one of the leading female actors in the discussion of Sweden's NATO application.

In Table 4.8, it is shown that the term *kamp*, for "struggle" or "fight" is associated more with a struggle for freedom (frihetskamp or *befrielsekamp*) than the textbook definition in the pre-training data, as this is specific to the context in which most of the Twitter users are posting. The term for freedom, *frihet* is associated with self-determination, independence, and sovereignty by the users in the Twitter data set. Note that there is a strong relationship between *kamp* and, for example, frihetskamp in the pre-training data, but it is not the closest relationship. What is observed here is that the similarity between *kamp* and concepts such as struggle for freedom or a resistance movement are being preserved or slightly accentuated while the word embedding is being pulled out of other contexts.

The term *alliansfritt* is included in the table. The most similar words were *neutralitet* and derivations thereof for all groups. Here the relationship between *alliansfritt* and *alliansl*ö*st* is shown to be a strong one for some of the groups. While the words are synonymous in meaning, there is a

	frihet	kamp	alliansfritt
Base Model	makt yttrandefrihet	dragkamp kraftmätning	säkerhetspolitiskt nato
Pre-Invasion	självständighet värdighet	dragkamp batalj	föregångsland parterland
Post-Invasion	suveränitet självständighet	frihetskamp motståndskamp	allianslöst självständigt
Pre-Application	självständighet religionsfrihet	dragkamp befrielsekamp	allianslöst valrike
Post-Application	frigörelse självbestämmanderätt	befrielsekamp frihetskamp	modeland kommunistskt
Leiden 0	jämlikhet värdighet	motståndskamp befrielsekamp	självständigt kärnvapenfritt
Leiden 1	överlevand suveränitet	motståndskamp befrielsekamp	alliansslöst modeland
Leiden 2	självständighet suveränitet	dragkamp befrielsekamp	alliansslöst kommunistskt
Leiden 3	yttrandefrihet jämlikhet	dragkamp kraftmätning	israelvänligt natoskydd
Infomap 0	självständighet suveränitet	befrielsekamp dragkamp	tvåspråkigt föregångsland
Infomap 1	jämlikhet sjävbestämmanderätt	motståndskamp befrielsekamp	tvåspråkigt självständigt
Infomap 2	yttrandefrihet åsiktsfrihet	dragkamp motståndskamp	alliansslöst tvåspråkigt
Infomap 3	makt yttrandefrihet	dragkamp kraftmätning	säkerhetspolitiskt demilitariserat

Table 4.8: Words with top cosine similarity to several keywords in the largest communities gathered from the Full Network setup for the CBOW model.

negative connotation to *allianslöst* with roughly the same distinction as when one would refer to someone living on the street as being ”homeless” rather than ”home free”, implying that an alliance is a good thing. The terms such as *tvåspråkigt* are related to the implication of non-alligned or neutral countries being multi-ethnic.

4.2.2 Skip-Gram vs. Continuous Bag of Words

The skip-gram models performed relatively poorly in finding a word embedding for *nato* and *sverige*. Frequently, the top related words for various groupings of the Twitter data were stop-words such as *och*, *inte*, or *att*. This result is somewhat expected based on how the data was gathered and the nature of the skip-gram model. With *nato* appearing so frequently, during training the skip-gram Word2Vec model will consider it to be one of the most likely words to form the context for nearly every other word, putting it in the same category as articles, conjunctions, and prepositions.

Table 4.9 shows the differences in word embeddings for the CBOW and skip-gram models. The CBOW model did not pick up the relationship between *krim* and the relatively uncommon word *annekteringen*, which appears between 20-60 times in each time step. The CBOW results show word embeddings with the strongest relationships with *annekteringen* are those roughly synony-

	annekteringen	valet	fred	2014
Pre-Invasion CBOW	ockupationen erövringen	riksdagsvalet presidentvalet	försoning harmoni	2013 2012
Post-Invasion CBOW	erövringen ockupationen	riksdagsvalet euvalet	försoning samexistens	2006 2008
Pre-Application CBOW	erövringen ockuptationen	riksdagsvalet kommunalvalet	avpsänning stabiliet	2008 2006
Post-Application CBOW	ockupationen erövringen	riksdagsvalet valdagen	försoning jämlikhet	2007 2008
Pre-Invasion SG	krim ukraina	parlamentsvalet omvalet	försoning krig	ukraina krim
Post-Invasion SG	krim invasionen	höst parlamentsvalet	samexistens varaktig	krim statskuppen
Pre-Application SG	erövringen krim	höst september	nedrustnng försoning	2008 krim
Post-Application SG	krim interventionen	höst parlamentsval	alliansfrihet samexistens	ukraina krim

Table 4.9: Comparison of most similar word embeddings for Bag of Words and Skip-Gram models.

mous to the textbook meaning of the term–conquest, invasion, occupation. The skip-gram model, on the other hand, captures a relationship between *krim* and the concept of annextation. Similarly, the CBOW model does not show a similiarity between the upcoming election, *valet*, and when it will be held, while this relationship appears in the skip-gram model. The election was in September, in the autumn (*höst*) season. The relationship between peace and disarmament or neutrality only appears in the results of the skip-gram model, as does the significance of the year 2014.

Other relationships appeared only in the skip-gram word embeddings. Terms for crazy (*galen*) and dictator (*diktator*) appeared together in the context of other words, so these word vectors had greater similarity in the skip-gram models and did not appear at all in the top 10 lists for any of the CBOW models.

Considering this, it is clear that it is worthwhile to consider the two models together to conclude anything about a particular group's vocabulary usage. Hence, the results in the following sections will include skip-gram model results alongside CBOW model results.

4.2.3 Lemmatized Text vs. Unlemmatized Text

Lemmatizing the text returned worse results than the raw text. Table 4.10 illustrates several places where relationships were either lost due to the lemmatization of the word, or erroneously created. Lemmatizing *allianlöst* to *allianlös* obscured the relationship between the former and the antonym *natoanslutet*. While *allianlöst* would appear in the same context as *natoanslutet*, the equally common term *allianlös* would not, preventing the model from learning this relationship. Note that this is only comparing a list of the top 10 closest matches; there is a cosine similiarity between every one of the 1 million+ terms in the model, so while it is possible that there exists a meaningful relationship between the lemmatized word vectors, that is not found readily nor is it as strong as other relationships.

Terms such as *krigsalliansen* or *försvarsalliansen* specifically refer to NATO for many groups, while the indefinite terms *krigsallians* or *försvarsallians* do not. The lemmatization of these terms creates a strong relationship between the lemma and NATO. While that may be of interest, and

word	Raw Text CBOW	Lemmatized CBOW	Raw Text SG	Lemmatized SG
allianslös	alliansfri	alliansfri	alliansfri	alliansfri
allianslöst	natoanslutet	—	alliansfritt	—
krigsallians	försvarsallians kärnvapensallians	nato försvarsallians	försvarsallians försvarsunion	krigsorganisation nato
krigsalliansen	nato försvarsalliansen	—	nato entiteten	—
förskola	förskolan grunskola	daghem fritidshem	dotters dagis	daghem dagis
förskolor	grundskolor	—	förskolorna	—
vild	häftig spontan	kakabaveh amineh kakabaveh	satirisk tam	amineh kakabaveh kakabaveh
vilde	kurd kakabaveh	—	politisk kakabaveh	

Table 4.10: A comparison of the most similar word embeddings for different models using Infomap Community 0 from the fully connected network.

possibly desired in some cases, it can also mean a loss of information as once the text is lemmatized, it would not be possible to determine that the definite form is being used specifically in reference to NATO.

Likeswise, the shared context between *dotters* and *förskola* is lost when all words which share the lemma *förskola* are considered together. The relationship is drawn from references to a particular quote by Left Party leader Nooshi Dadgostar who said that she ”did not want nuclear weapons near her daughter’s preschool.”[N24] Both *dotters* and *förskola* appear together in the context for words such as *kärnvapen*. The term *kärnvapenbaser* is one of the top 10 most similar words to *förskolor* (but not in the top 2, so not shown in Table 4.10), but a similar relationship is not found with the lemmatized text. A correlation between a specific form or inflection of a word and an unrelated term that arises from a quote or similar is likely to be lost among other correlations between every other inflection of the word’s lemma and the terms that they appear with.

An erroneous relationship is created in the case of the lemma *vild*, meaning ”wild” or ”savage”. The term *vilde* can refer to those who live outside the bounds of civilization, but in the Swedish political context it means an independent member of the Riksdag without a party affiliation. The word in this context does not have a connotation of being ”wild” or ”savage”, but the lemmatization creates an association between Amineh Kakabaveh and the word *vild* which can give the wrong indication of the word’s intent.

Based upon these results, the lemmatized text was not considered in these results unless otherwise stated.

4.2.4 Top Level Communities vs. Subcommunities in Hierarchical Leiden

The Hierarchical Leiden Algorithm returned a large number of subcommunities for which the word embeddings do not show a clear distinction. When analyzing the data, it was important to understand what could be learned by looking at these subcommunities.

Figure 4.11 shows an example of the words most closely associated with the word *sossar*, slang for members of the Social Democrats party in Sweden. The word appears over a thousand times in both Community 1 and Community 2 of the Leiden partition on the full network.

The results are as expected across the board with only some exceptions. The term is most closely associated with other political parties. However, community 1,1 associates the Social Democrats with the term *fjortisar*, a derogatory term for immature teenagers. This relationship appears in the top level community. The slur that community 2,3 associates with the Social Democrats does not appear in any of the other communities, and given that the top level community is 3.5 times as large as subcommunity 2,3, not often enough at the top level to appear in the list of most closely associated words there.

From this and other similar cases, it can be seen that similarity between word embeddings within a community partitioned by the Leiden Algorithm may only reflect the usage among a small subset of that community, and that some awareness of the heterogeneousness of the subcommunities within each community is needed to avoid drawing broad conclusions about the language user of the larger community. It is also shown that some strong opinions are held by groups too small to influence the word embedding of the larger community, so that echo chambers and words of interest are more likely to exist within the subcommunities.

This is illustrated by Figure 4.5, which shows the cosine similarity between selected terms for community 0 and 2 from the Leiden algorithm's partition of the full network. The subcommunities are placed somewhat more expectedly–anti-NATO communities find a connection between Azov and Nazis, while pro-NATO communities have a stronger association between freedom and security. This predictable placement is lost when the subcommunities here are assembled into the larger supercommunities, and the relationship between the subcommunity's average position and their use of language is obscured. It is possible that other outliers, such as the weakly pro-NATO community with the highest cosine similarity between *frihet* and *säkerhet* in the figure, are also composed of a mix of pro- and anti-NATO sub-subcommunities which, if separated out, would have placement following a similar pattern. This is difficult to know, as the data from the community in question quickly becomes too small to use to train a Word2Vec model with as it is repeatedly broken up.

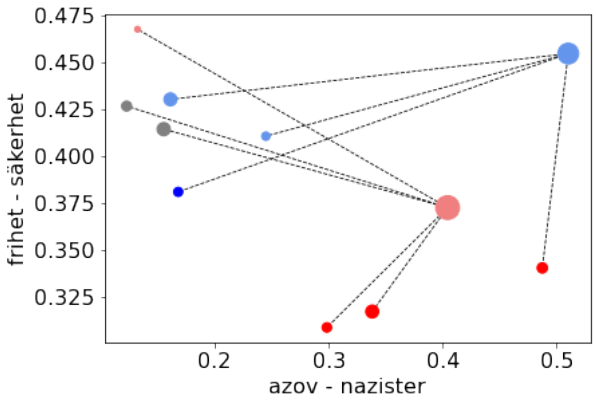

Figure 4.5: The distribution of communities in a space defined by the cosine similarity between *frihet* - *säkerhet* and *azov* - *nazister* according to the CBOW model for communities 0 (weakly anti-NATO in light red) and 2 (weakly pro-NATO in light blue) from the Leiden algorithm run on the full network. The dotted lines shown the placement of the subcommunities within the same space.

	Leiden Community 1 - Full		Leiden Community 2 - Full	
	sossar CBOW	sossar SG	sossar CBOW	sossar SG
Top Level	socialdemokrater folkpartister vänsterpartister	socialdemokrater fjortisar pkiter	socialdemokrater moderater vänsterpartister	kommunister socialister s
Subcommunity 0	socialdemokrater folkpartister moderater	högersympatisörer folkpartister centerpartister	moderater socialdemokrater vänstepartister	folkpartister moderater kristdemokrater
Subcommunity 1	socialdemokrater moderater vänsterpartister	socialdemokrater fjortisar kärringar	moderater socialdemokrater miljöpartister	moderater socialdemokrater folkpartister
Subcommunity 2	socialdemokrater moderater vänsterpartister	riksdagsmän feminister invandringskritiker	socialdemokrater centerpartister moderater	socialdemokrater folkpartister s
Subcommunity 3	moderater socialdemokrater vänsterpartister	moderater vänsterpartister socialdemokrater	moderater socialdemokrater folkpartister	fittor kristdemokrater er

Table 4.11: An example comparing the top-level community results with the subcommunity results for the term *sossar*.

4.2.5 Time-based Changes

Tables 4.12 shows a selection of words and the top 2-4 words by cosine similarity for the CBOW models. The progression of topics follows the developments of Sweden's NATO application.

Once the NATO application was submitted, there was an increased association between *natoansökan* and the *natoprocess*. As the application became a common topic of conversation, a hashtag become common enough to be picked up as a similar term in the dataset. Also of note is that once Turkey had made their objections in the pre-application period, an association emerged between Turkish President Recep Tayyip Erdogan and both Vladimir Putin and the PKK. In the post-invasion period, an association between NATO opponents and loyalty to Putin is seen.

Tables 4.13 shows the same for the SG results. Here, the connection between *destabilisera* and the *natoansökan* is observed. Tweets using terms related to destabilization appeared almost exclusively in the post-application period. The cosine similarity between defense (*försvaret*), and word embeddings related to the 2% of GDP spending goal for NATO members emerged in the post-application period as the topic of defense budget changes became relevant. In the post-application period, when questions of Turkish human rights records were loudest, *erdogan* and *diktator* began to have close similarities.

Other-time related word embedding relationships are found. Users associated *inkompetent* with embattled Justice Minister Morgan Johansson during the Vote of No-confidence. Resistance to NATO began to be associated more strongly with the Green Party and individual Social Democrats such as Peter Hultqvist rather than with the Social Democrats as a whole after the NATO application was submitted.

The use of the romanization of the Ukrainian pronunciation of the capital city of Ukraine, *kyiv* becomes popular only after the invasion of Ukraine. The use is most common in the post-invasion period, and becomes used much less after the NATO application was submitted. We do see an association with *groznyj*, the capital of Chechnia which was devastated by Russia in the Second

Word	Pre-Invasion	Post-Invasion	Pre-Application	Post-Application
natoansökan	medlemsansökan osansökan natooption medlemskaps- ansökan	natoanslutning anslutning dispens- ansökan medlemskaps- ansökan	medlemskaps- ansökan natoanslutning #natoansökan anslutning	natoanslutning natoprocess(en) nato- medlemskapet natoanslutning
erdogan	mubarak sarkozy gül sharon	turkiet orban mubarak berlusconi	turkiet putin turkarna pkk	turkiet putin turkarna #erodgan
vilde	kungamakare kristdemokrat vågmästare	krisdemokrat rabulist vågmästare	kungamakare partimedlem inkryssad	vänstervilde kurd kakabaveh
kiev	tblisi lviv	kyiv charkvi	istanbul budapest	donetsk leningrad
nato- moståndare	invandrings- kritiker emumotståndare vänster- intellektuella	natoförespråkare putinister natoanhängare	natoförespråkare natoanhängare sossar	miljöpartister kärnkrafts- motståndare skeptiker

Table 4.12: Words with top cosine similarity for the CBOW models grouped by date. Words that appeared with similar placement in all time steps are omitted here.

Chechen War.

Figures 4.6 - 4.12 are examples of the changes in cosine similarity between given keywords and other related terms over time. In Figure 4.6, the growing association between Amineh Kakabaveh and the *vilde* status over time outpaces the earlier association of the term with being a sort of political gatekeeper due to a lack of party affiliation. Likewise, Figure 4.7 shows the strong association between Sweden and NATO that is unique to this dataset, as compared with the pre-trained model, as well as a growing trend to hold NATO synonymous with *krigsallians*, a term which hardly appears in the pre-training corpus.

Figure 4.8 shows the rise in the use of *kyiv* to refer to the capital of Ukraine. Even though the absolute cosine similarities with terms like *tbilisi* and *moskva* remain relatively unchanged (and so were omitted in Table 4.12 due to similarity across time periods), the emergence of a new associated term represents an interesting change in the embedding for *kiev*.

Figures 4.9 - 4.10 show the changing usage of the term *natoansökan*. The CBOW model reveals how the term *ansökan* becomes almost synonymous with *natoansökan*, as that is the only application which users are refering to. In both SG and CBOW models we see the term *natoanslutningen* become associated with the application during the build up to the application being filed as the conversation moves on to the accession process. The ”-en” suffix refers to the definite form of this term in Swedish, thus ”the NATO accession”, which refers specifically to Sweden’s NATO accession, a prospect considered unthinkable in the first two time steps.

Peaks are seen in Figure 4.11 and 4.12 for the association between preschools and nuclear weapons after Dadgostar’s comments, and the sudden association between *minister* and both *inkompetent* and *regeringskris* comes as the vote of no confidence moves forward. With greater resolution, and more data, this type of analysis could see the users’ change of opinion of certain topics by

Word	Pre-Invasion	Post-Invasion	Pre-Application	Post-Application
natoansökan	medlemsansökan byggför- handlingarna drömregering	medlemsansökan natomedlemskap destabilisera(nde)	natoanslutning eventuell natomedlemskap	sveriges finlands natoprocessen
försvaret	försvarsmakten invasionförsvaret förbandsverksamhet	försvarsmakten bnp rusta	bnp 2% rusta	bnp 2% försvarsanslaget
erdogan	orban recep janukovitj	orban islamisten trump	kurder pkk putin	putin diktator(n) bashar al assad
kiev	donbass teheran	kyiv damaskus	abchazien kyiv	groznyj smolensk
nato- moståndare	vänsterpartisterna usahatare realisterna	–	flyktingkramare usahatare natoförespråkare	hultqvist natoförespråkare kärnkrafts- motståndare

Table 4.13: Words with top cosine similarity for the Skip-grame models grouped by date. Words that appeared with similar placement in all time steps are omitted here.

looking for these sharp changes.

4.2.6 Pro- and Anti-NATO Rated Communities

In general, the vast majority of a priori determined words of interest which appeared in sufficient numbers in both groupings showed no clear indication of new or different usage patterns. Terms such as *nato*, *försvar*, *neutralitet*, and other common words provided no indication of any difference between the groups.

The automated method for finding words of interest yielded several hundred potential candidates consisting of clear indicators of a position on the NATO question, non-obvious relationships which proved interesting, and then many junk terms: names of infrequently mentioned politicians, obscure emojis, common misspellings, English words, slang or unusual abbreviations, or other words which when evaluated showed little of interest.

Words with obvious pro- or anti-NATO leanings, for example *#wearenato*, *natopropaganda*, *natoälskare*, *#nejtillnato*, *#nejtilleu*, *#swexit*, *krigsmaskin*, etc., were used relatively infrequently overall. Many words associated with the topic of NATO membership, such as *finlandisering*, *ansökningsperioden*, or *bombdåd*, did not have novel uses among the different user communities when the word embeddings were evaluated.

A comparison of word embeddings from Pro- and Anti-NATO Communities is listed in Tables 4.5 and 4.6. For each word in the table, the top 10 most similar words of each community in the set were collected, and the highest rated novel words (unique to that set and not appearing in the top 10 most similar words of the other) from any of the communities in the set are listed along with the cosine similarity between the novel term and the input term. These words are among those found by the methods previously mentioned.

Words which only appeared frequently in one of the types of communities (either anti- or pro-NATO) showed some difference, though the novel terms did not appear in all subcommunities. For example, terms related to Azov Battalion, a paramilitary group in Ukraine, such as *azov* and

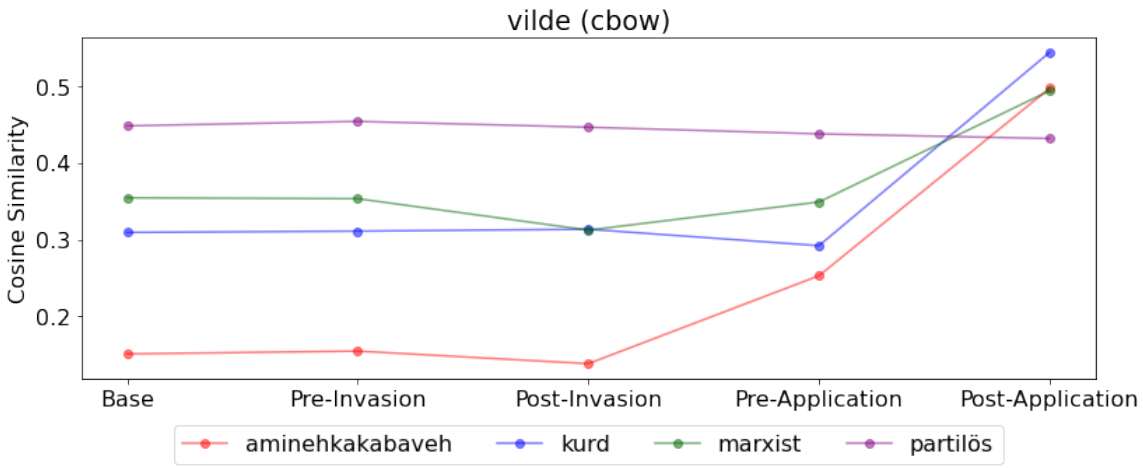

Figure 4.6: The change in cosine similarity between *vilde* and select terms over time.

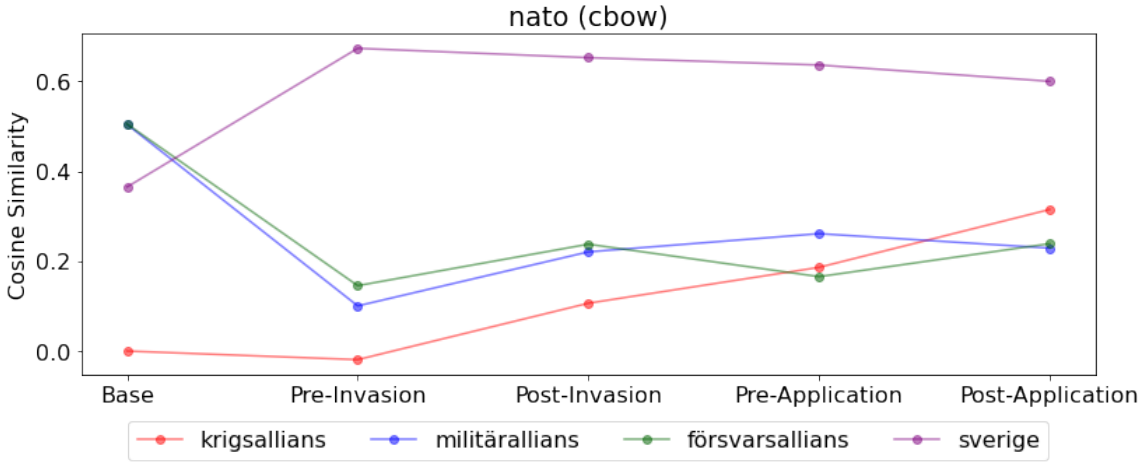

Figure 4.7: The change in cosine similarity between *nato* and select terms over time.

azovbataljonen appeared 516 among the pro-NATO groups and 2179 times among the anti-NATO groups. Nearly all of the anti-NATO leaning groups had a term such as *nazist*, *skinheads*, or *nazistiska* as the most closely-related term to these terms, while no Nazi-relationship appeared in the top 10 closest words among any of the pro-NATO groups.

The term *vilden* or *vilde*, again referring to a member of the Swedish Riksdag without a political party affiliation is used much more often by pro-NATO groups, with 2963 appearances in pro-NATO groups and 232 in anti-NATO groups.

Terms *allianslöst* and *allianslös* do not appear in sufficient numbers in the anti-NATO groups to have a word embedding, since the Word2Vec model ignores words that appear less than 10 times. These terms appear more than 1100 times in the pro-NATO groups considered here.

NATO proponents often mocked NATO opponents by referencing Dadgostar's quote about nuclear weapons and preschools. The association between nuclear weapons and preschools (*förskolor*) was picked up by the skip-gram Word2Vec model, as the words would appear together in the context of other terms. Similarly, only pro-NATO communities had a relationship between *förskola* and *dotters*. This association was only seen in the top matches for pro-NATO communities. Anti-NATO communities rarely used the terms *förksolor* or *förksola* and no relationships were observed other than those inherited from the base model.

Terms such as *wef* and *tigray* showed a high degree of multimodularity. Mentions of Tigray

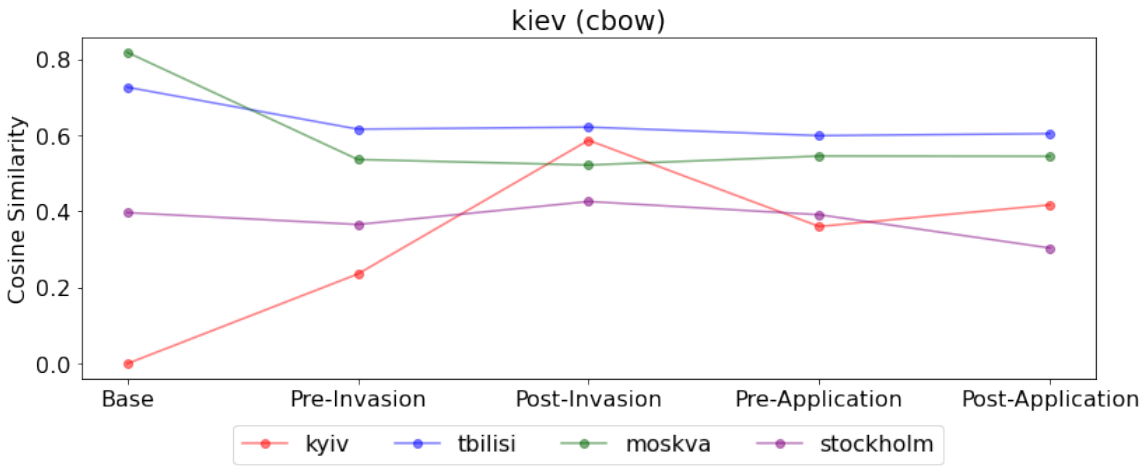

Figure 4.8: The change in cosine similarity between *kiev* and select terms over time.

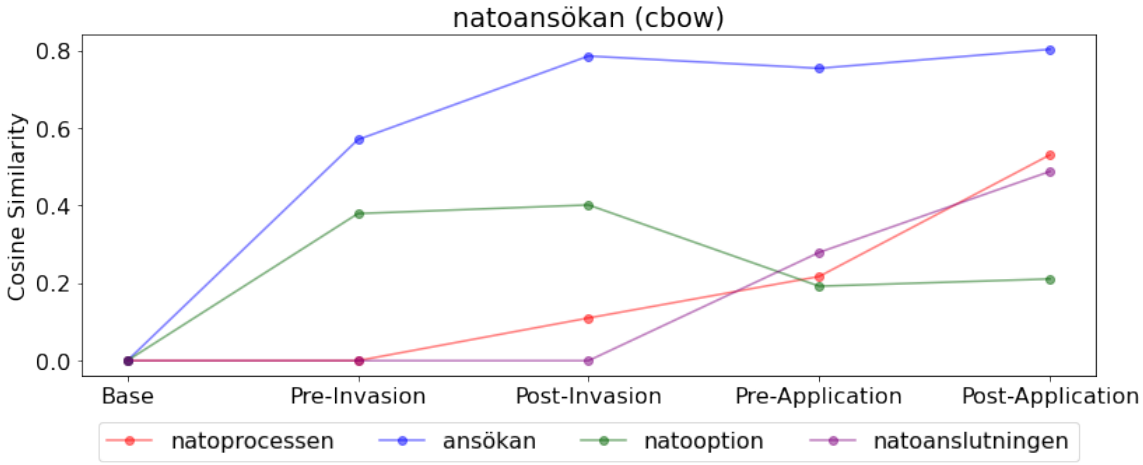

Figure 4.9: The change in cosine similarity between *natoansökan* and select terms over time.

were almost entirely contained within Community 6 of the Leiden Algorithm's partition of the full network, and quick inspection showed that most of the discussion in the community's small number of tweets was about the war in Ethiopia. Mention of the WEF, the World Economic Forum, was found to be far more common in those communities evaluated as anti-NATO, and word embedding comparison showed that the term was closely related to various conspiracy theories in those communities. Word vectors for *wef* were more closely related to conspiracy-theorist terms such as *nwo*, for the "New World Order", in the anti-NATO group, though they did not appear in all of the subgroups. For a quick comparison in usage, the term *nwo* was used 1138 times in the anti-NATO communities, and only 76 times in the pro-NATO communities, while *wef* appears 3933 times in the anti-NATO communities and 448 times in the pro-NATO communities.

Pro-NATO communities showed the term *annekteringen* as one of the most similar words to the keyword *2014*, while anti-NATO communities saw *statskuppen* as one of the most similar words. Word relationships which showed that *aggressiv* and *granne* appeared together often in the context for other terms were also limited to the pro-NATO communities.

Mention of the *vita bussarna* appear in the pro-NATO communities. The white buses were used to evacuate (primarily Scandinavian) prisoners from German concentration camps during the Second World War. Other aspects of the Second World War were raised by the pro-NATO communities, including a relationship between *chamberlain* and *adolf* which arises from tweets such

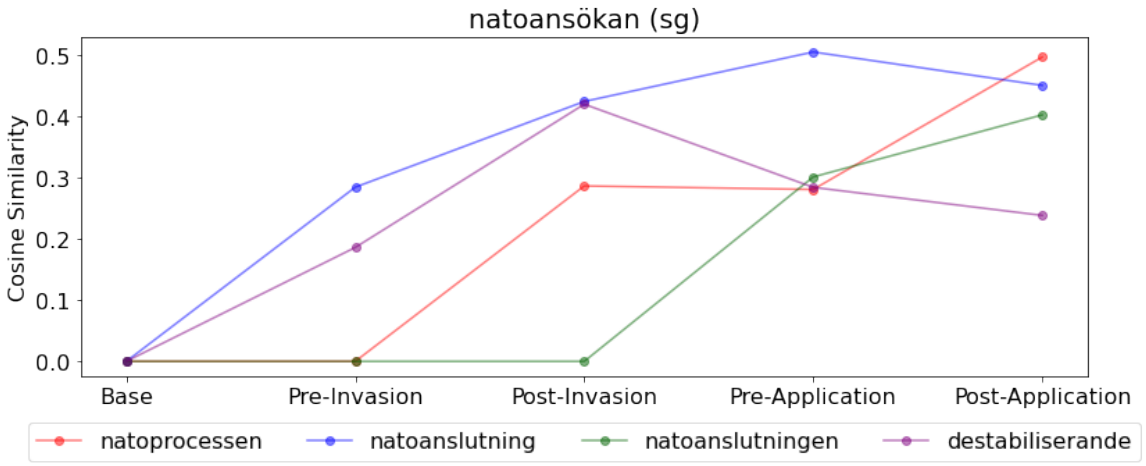

Figure 4.10: The change in cosine similarity between *natoansökan* and select terms over time.

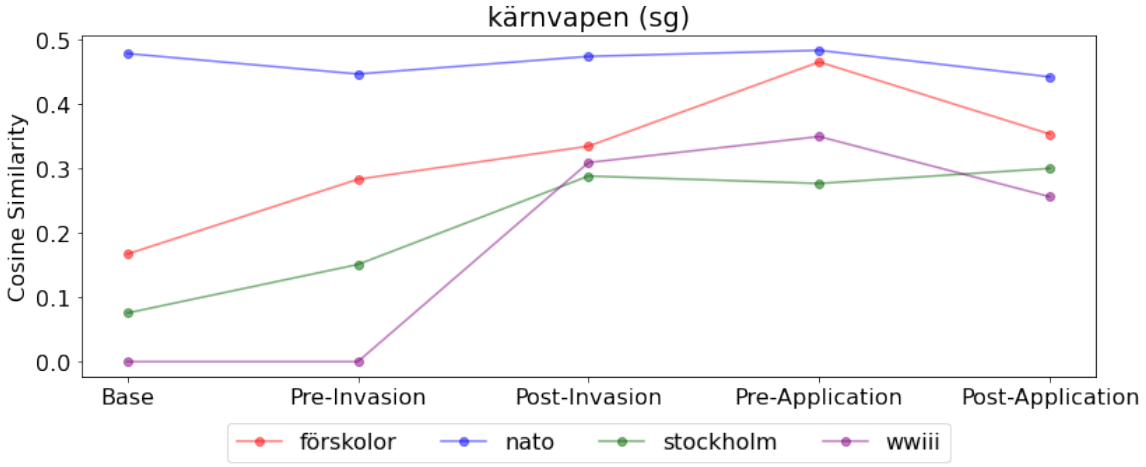

Figure 4.11: The change in cosine similarity between *kärnvapen* and select terms over time.

as:

> Fråga dig själv. Om Nato funnits 1938 med motsvarande militära styrka & solidaritet. Hade då Adolf annekterat Österrike, hade pacifisten Chamberlain (Peace in our time) då givit bort Sudetenland till Adolf, Hade Adolf vågat angripa Polen 1939 & startat WW2?

An important consideration is that these groups were determined by choosing groups which showed the strongest position indication as a result of manual annotation. The groups did not, in general, standout from looking at the full set of all data. For example, the comparison of Crimea to Kosovo was presumed to be a strong anti-NATO indication, however it also appears in Community 2 of the Leiden Full Network, which is strongly pro-NATO rated. (Of the subcommunities of Community 2 of the Leiden Full Network, only 2,1 showed *kosovo* as one of the top 10 most similar words to *krim*.) Looking at the top 10 most similar words to a set of keyword for this community would give indications of both pro- and anti-NATO leanings, thus leaving no real conclusion.

Word	Pretrained Model	Pro-NATO Communities	Anti-NATO Communities
putin	medvedev bush	putler hitler	zelensky trump
azov	tver belgorod	rostovnadonu ukrainasfotballs-landslag	bataljonen nazister
kakabaveh	miriam kohandlat	vilden morgan	kurdiska vänsterpartiet
de	dom dem	sossarna —	— —
morgan johansson	maria larsson folkhälsominister	mogge regeringskris	berit andnor s regeringen
ingå	inkludera delta	— —	militärallians krigsallians
krim	krimhalvön kiev	annekteringen georgien	kosovo nagornokarabach
förskolor	skolor grundskolor	kärnvapen kärnvapensilos	— —
förskola	skola grundskola	dotters kärnvapen	— —
wef	worldeconomicforum oecd	— —	nwo who
adolf	otto gustaf	chamberlain sudentenland	gestapo —
alliansfritt	neutralt säkerhetspolitiskt	allianslöst —	— —
aggressiv	våldsam hotfull	kärnvapendiktatur granne	— —
2014	2013 2015	annekteringen —	statskuppen kuppen
vita	svarta röda	bussarna —	— —

Table 4.14: A small selection of word cosine similarities from pro- and anti-NATO communities. Words that are common to both pro- and anti-NATO communities are omitted, and not all communities with the same rating included the same unique words.

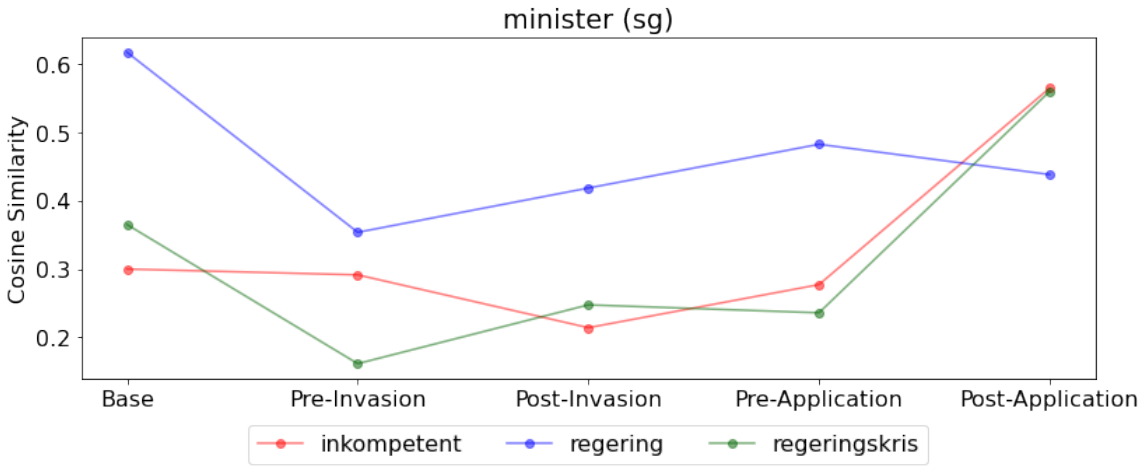

Figure 4.12: The change in cosine similarity between *minister* and select terms over time.

4.2.7 Community Position

To investigate whether or not it would have been possible to discover ideologically-identifying terminology without annotation, an analysis of the clustering of the communities within a 2D space defined by the cosine similarity of two pairs of words was conducted. Several examples of the results are given in Figures 4.13 - 4.21. In the figures, the communities are shaded by their NATO score to give an indication of how likely pro- or anti-NATO clusters which had the properties of those discussed in the previous section could have been found without annotation.

Generally, any relationship between the supposedly identifying terminology is weak. Strongly pro-NATO communities for which one of the words of interest is within the top 5 ”most similar words” to a given word are positioned in the space beside strongly anti-NATO communities for which the cosine similarity between the word pair is not even with the top 10 ”most similar” for the given word. Some general trends can be observed, however in none of the figures is there a distinct clustering of communities which one can say with certainty are clearly ideologically similar.

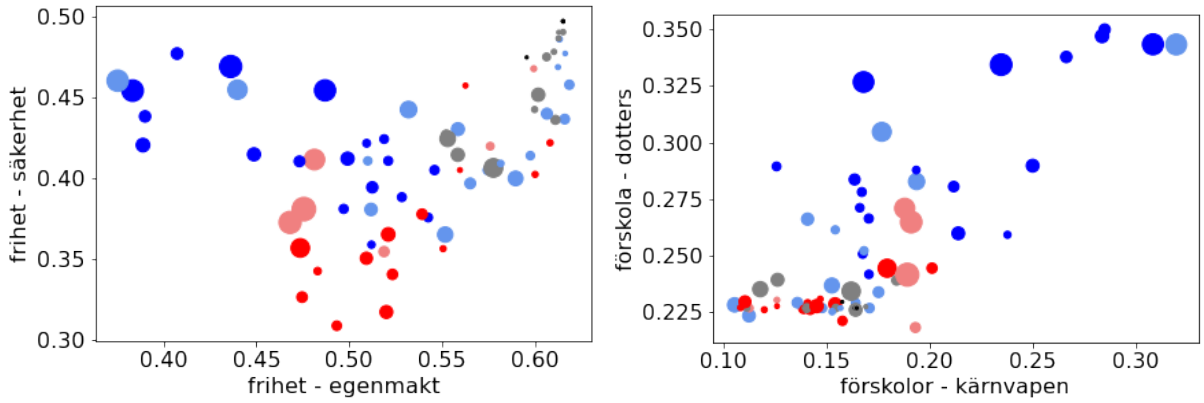

Figure 4.13: The distribution of communities in a space defined by the cosine similarity of the terms on the X- and Y-axes according to the CBOW model for each community found from any of algorithms. The coloration shows the average NATO score of the community from the hand annotation and the size is proportional to the community size.

For example, in Figures 4.13, the relationship between *dotters* and *förskola* as well as *förskolor* and *kärnvapen* is shown to be stronger with pro-NATO communities in general, however there are

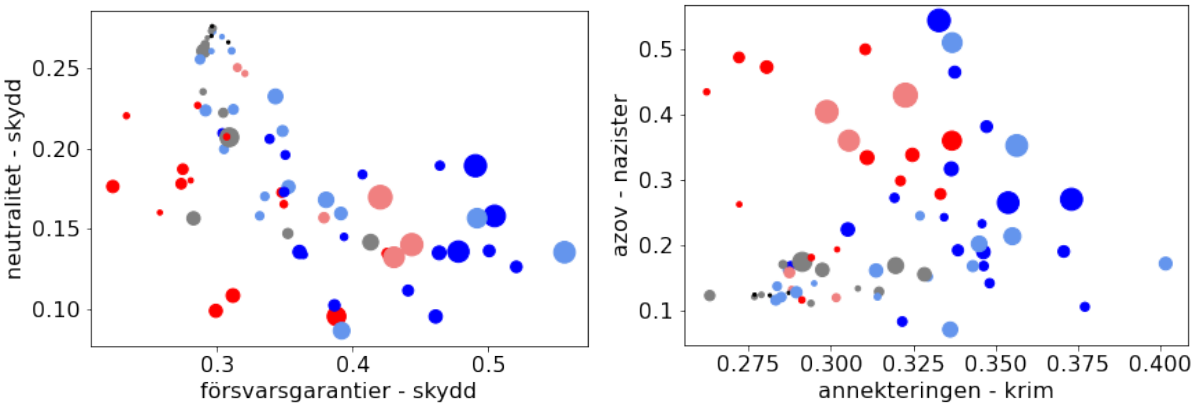

Figure 4.14: The distribution of communities in a space defined by the cosine similarity of the terms on the X- and Y-axes according to the CBOW model for each community found from any of algorithms. The coloration shows the average NATO score of the community from the hand annotation and the size is proportional to the community size.

many pro-NATO communities which do not make such references and many pro-NATO communities for whom the relationship is weaker than in anti-NATO communities. In the same figure, a trend towards equating freedom with security is seen to be stronger in pro-NATO communities.

In Figure 4.15, a greater proclivity to link NATO with a *krigsallians* is seen among the anti-NATO communities, while the pro-NATO communities would have a stronger relationship with *försvarsallians*, indicating the role that each group of communities believes that NATO plays in the world. Also, this gives some indication that, for anti-NATO users, NATO is more synonymous with the United States of America, criticisms of which are often used as reasons against Sweden joining NATO. The outlier here is from the Hierarchical Leiden algorithm. Community 3,2 from the likes only network has a strong association between *usa* and *nato* with a cosine similarity greater than 0.7. This community also exhibits other trends which indicate that they are strongly anti-NATO: *2014* has *statskuppen* as its most similar term when the Skip-gram model is used, and with the CBOW model finds *nazister* as the most similar word to *azov* and *krigshetsare* most strongly connected to *natoälskare*, suggesting that this community, as a whole, sees NATO supporters as war-mongers and is less sympathetic with Ukraine. That this outlier could be identified with these methods and that the community possessed other semantic evidence of the user position stands as evidence that there is some truth to the hypothesis presented as the research questions for this work.

Figure 4.18 gives an indication of political leanings more than a position on NATO. Naturally, those communities which are most likely to liken the Social Democrats to irresponsible teeenagers are also more likely to link the Left Party with communism. The use of the term *vilde* appears more often with those communities who are pro-NATO, likely due to the larger number of complaints about the arrangement between Kakabaveh and the Social Democrats during the Vote of No Confidence, but there is no strong relationship between pro- and anti-NATO position and the relationship between *vilde* and *kakabaveh*. Connections between *marxist* and *vilde* were stronger among pro-NATO rated groups, but even given that relationship it is not possible to draw a clear line between pro- and anti-NATO leanings.

Figure 4.19 illustrates the impossibility of clustering the communities based upon the relationship between *krim* and other conflict areas as communities fill a broad spectrum of cosine similarity values between these terms. In this same figure, the association between Turkey and various forms of the word for dictator verifies that relationship between anti-NATO sentiment and criticism of Erdogan.

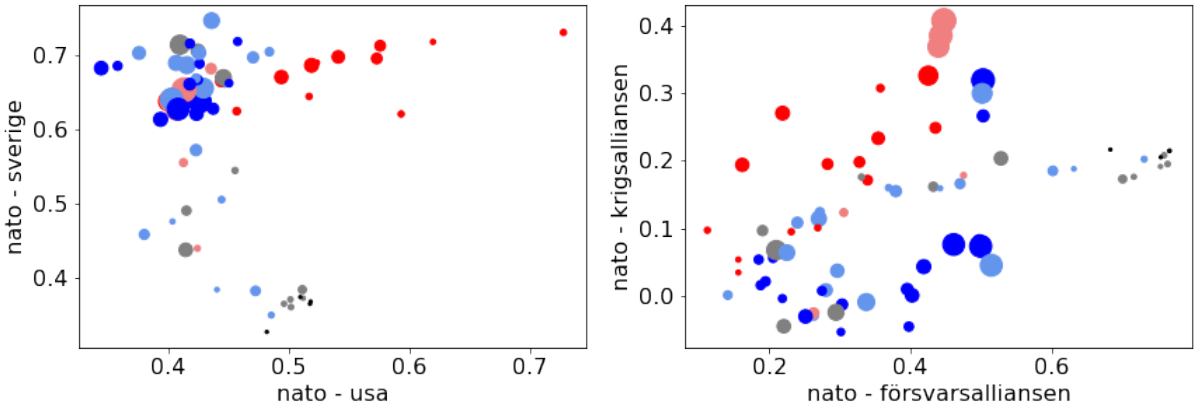

Figure 4.15: The distribution of communities in a space defined by the cosine similarity of the terms on the X- and Y-axes according to the CBOW model for each community found from any of algorithms. The coloration shows the average NATO score of the community from the hand annotation and the size is proportional to the community size.

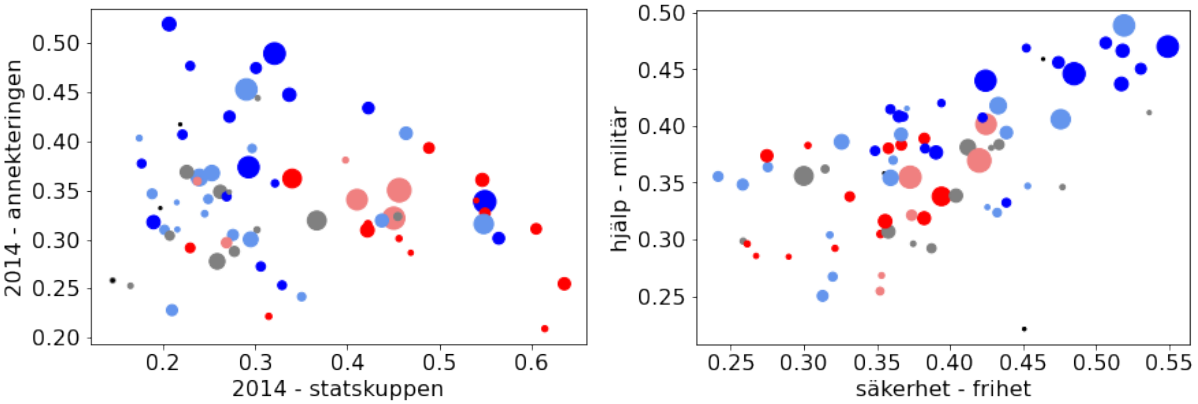

Figure 4.16: The distribution of communities in a space defined by the cosine similarity of the terms on the X- and Y-axes according to the SG model for each community found from any of algorithms. The coloration shows the average NATO score of the community from the hand annotation and the size is proportional to the community size.

Figure 4.20 shows a divide in NATO opinion based upon whether *inkompetent* appears in the same context as *minister*, a nod to the vote of no confidence for Morgan Johansson and the disdain for him from those on the political right. A relatively strong indicators of group membership is in the cosine similarity between *stark* and *ensam* in Figure 4.20. Counter-intuitively, the pro-NATO groups showed a larger cosine similarity between these terms. As previously mentioned, cosine similarity does not always indicate that words are synonyms, just that they appear in the same context. The conceptual relationship between these two terms is established by the pro-NATO users making an argument that Sweden is not strong if it is alone; the anti-NATO side is not attempting to make an argument that being alone is being stronger.

Finally, in Figure 4.21 a slightly greater similarity between *putin* and *psykopat* as well as a considerably stronger association between *bucha* and *blodbad* illustrates a common argument for the pro-NATO posters: cities and towns in Sweden could share the fate of Bucha, a city in Ukraine on the outskirts of Kiev where the Russian Federation has been alleged to have carried out a massacre of civilians. This threat, many of these users argue, arises from an irrational foreign policy conducted by Vladimir Putin. This echos the association between Vladimir Putin and Adolf Hitler which

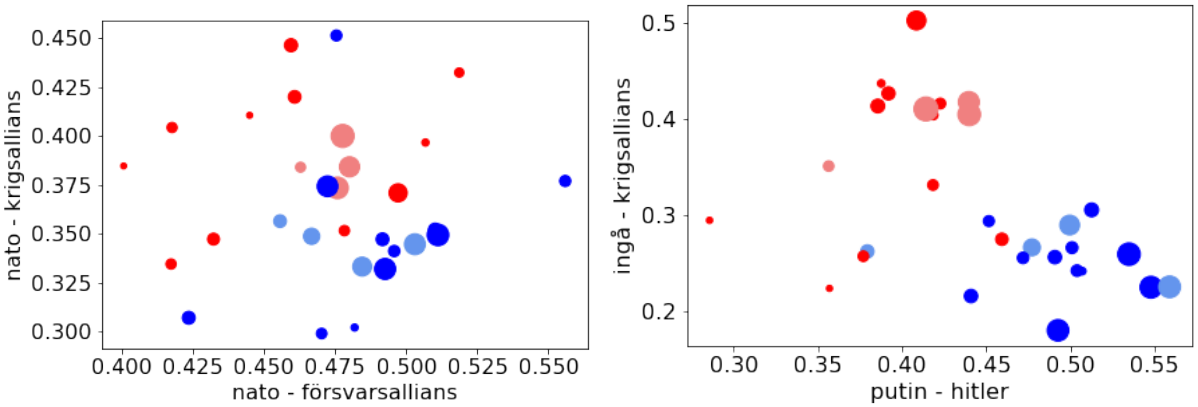

Figure 4.17: The distribution of communities in a space defined by the cosine similarity of the terms on the X- and Y-axes according to the SG model for each community found from any of algorithms. The coloration shows the average NATO score of the community from the hand annotation and the size is proportional to the community size.

appears often among pro-NATO communities.

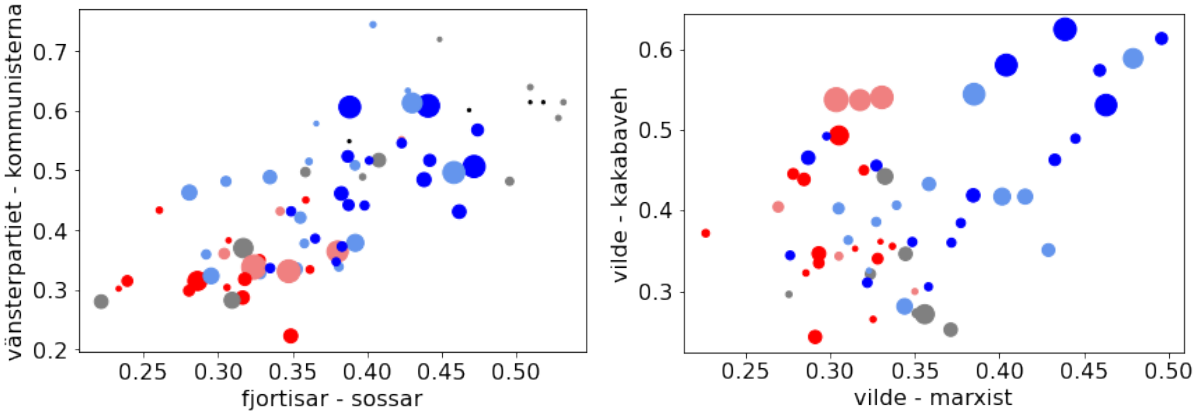

Figure 4.18: The distribution of communities in a space defined by the cosine similarity of the terms on the X- and Y-axes according to the SG model for each community found from any of algorithms. The coloration shows the average NATO score of the community from the hand annotation and the size is proportional to the community size.

4.2.8 Pro- and Anti-NATO Users

A final assessment was done of all users grouped by their NATO scoring, ignoring the communities found through the community detection algorithm. The idea here was to directly compare ideological positions to determine if there were clear differences in language that were missed when splitting the groups up into communities. Figure 4.22 shows the relative size of these groups, as well as the degree of connection between them for all types of connections except likes. Users with similar positions do follow, retweet, and quote each other more frequently. The anti-NATO users are somewhat more tightly connected, preferring to follow other anti-NATO users at a 6:1 ratio vs. pro-NATO users, who have a 3:1 ratio favoring their side. What is notable here is that the connections between strongly pro-NATO and strongly anti-NATO groups is greater than either groups interaction with the large group of neutral users, or between the weakly pro-NATO and weakly

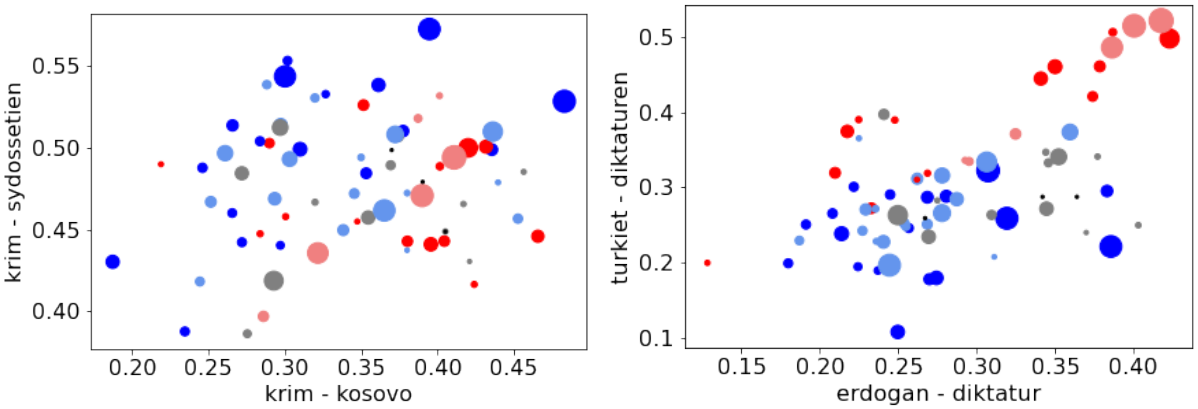

Figure 4.19: The distribution of communities in a space defined by the cosine similarity of the terms on the X- and Y-axes according to the SG model for each community found from any of algorithms. The coloration shows the average NATO score of the community from the hand annotation and the size is proportional to the community size.

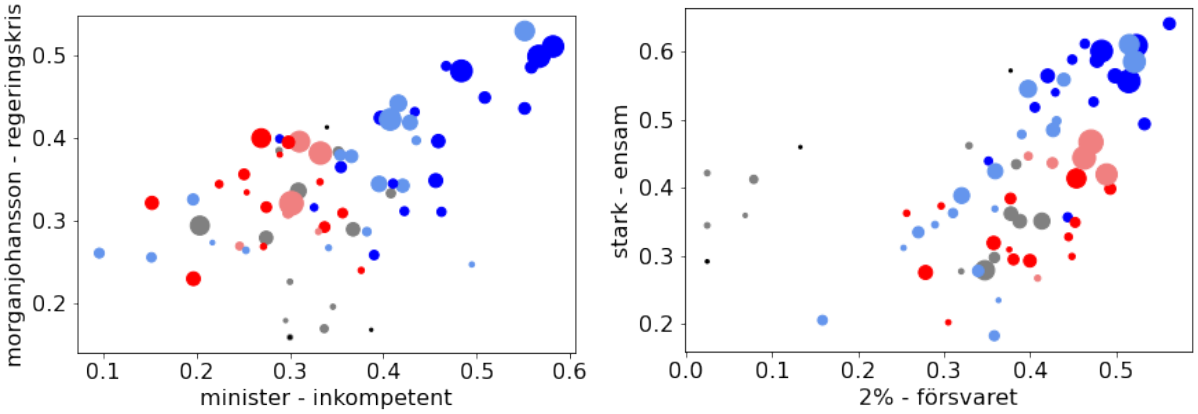

Figure 4.20: The distribution of communities in a space defined by the cosine similarity of the terms on the X- and Y-axes according to the SG model for each community found from any of algorithms. The coloration shows the average NATO score of the community from the hand annotation and the size is proportional to the community size.

anti-NATO groups.

Table 4.15 gives some of the cosine similarities between words of interest from a corpus of users separated by their position. The observations track closely with those from the previous section.

Referring back to the changes in cosine similarity over time, it is also advantageous to look at how the language use of each side of the NATO debate changes with time. Figures 4.23 - 4.25 illustrate the differences in word usage among the pro- and anti-NATO rated users for a selection of words of interest.

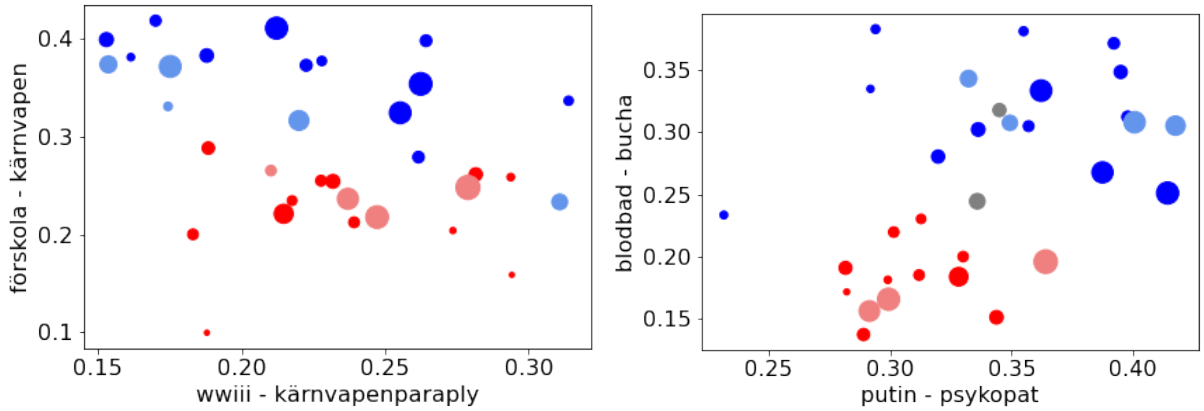

Figure 4.21: The distribution of communities in a space defined by the cosine similarity of the terms on the X- and Y-axes according to the SG model for each community found from any of algorithms. The coloration shows the average NATO score of the community from the hand annotation and the size is proportional to the community size.

Word Pair	Pro-NATO	Anti-NATO	Neutral
propaganda – natopropaganda (CBOW)	0.34	0.47	0.33
kiev – kyiv (CBOW)	0.64	0.38	0.33
azov – nazister (SG)	0.39	0.59	0.45
2014 – statskupp (SG)	0.23	0.50	0.32
dotters – förskola (SG)	0.67	0.42	0.48
frihet – säkerhet (CBOW)	0.50	0.36	0.38
militärallians – krigsallians (CBOW)	0.39	0.64	0.43
nato – krigsallians (SG)	0.32	0.41	0.28
kakabaveh – vilde (CBOW)	0.64	0.32	0.42
minister – inkompetent (SG)	0.51	0.32	0.42
nwo – wef (SG)	0.37	0.63	0.43
putin – putler (CBOW)	0.72	0.31	0.33

Table 4.15: A selection of cosine similarities between words of interest for a corpus composed of all tweets by users rated pro-NATO, anti-NATO, or neutral.

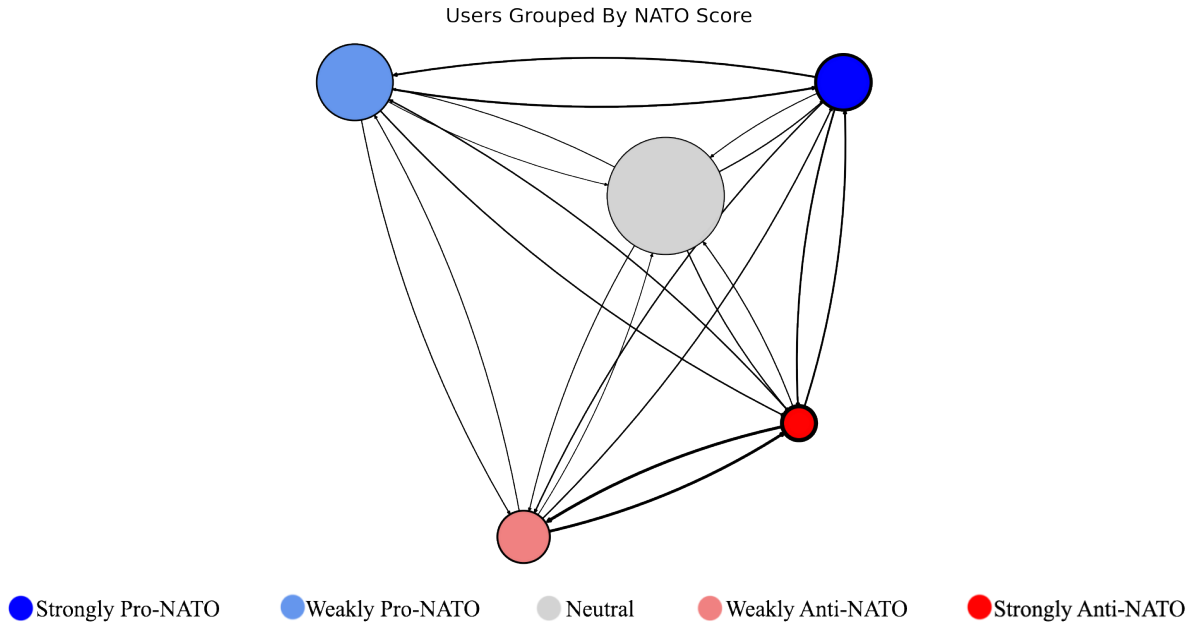

Figure 4.22: Groups of users of a given NATO position. The size of the nodes indicates the size of the group, and the thickness of the connections is proportional the probability to follow, retweet, or reply to a user in another group. The thickness of the node outline indicates the in-group probability.

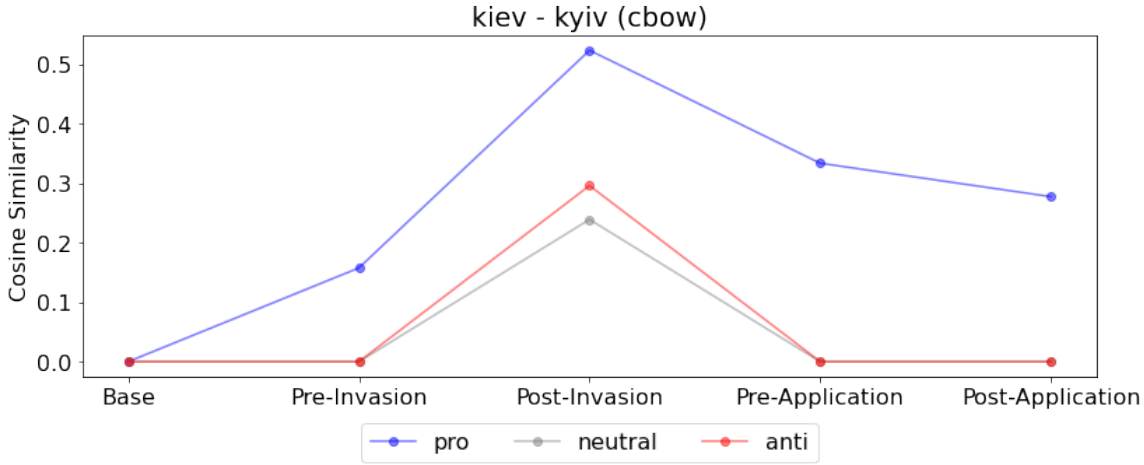

Figure 4.23: The change in cosine similarity between *kiev* and *kyiv* over time for pro-NATO and anti-NATO users.

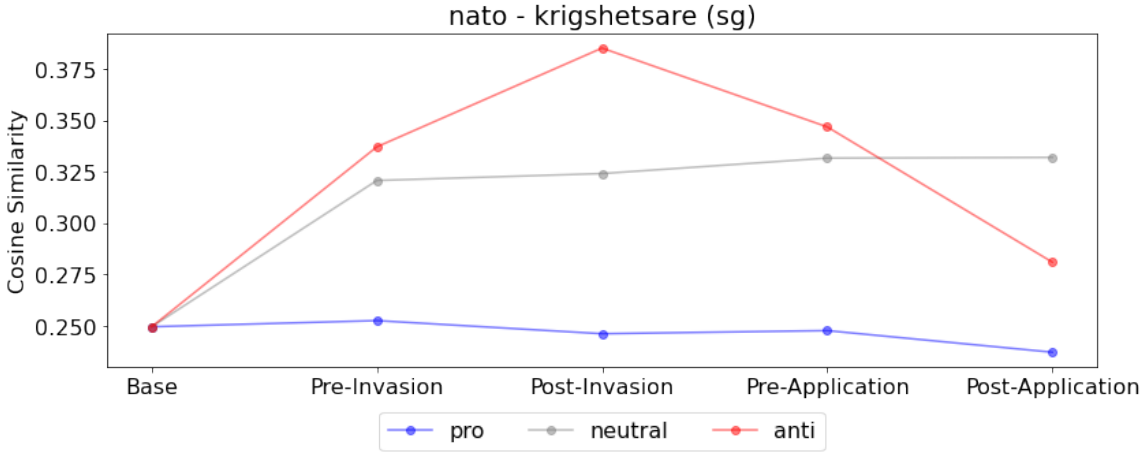

Figure 4.24: The change in cosine similarity between *nato* and *krigshetsare* over time for pro-NATO and anti-NATO users.

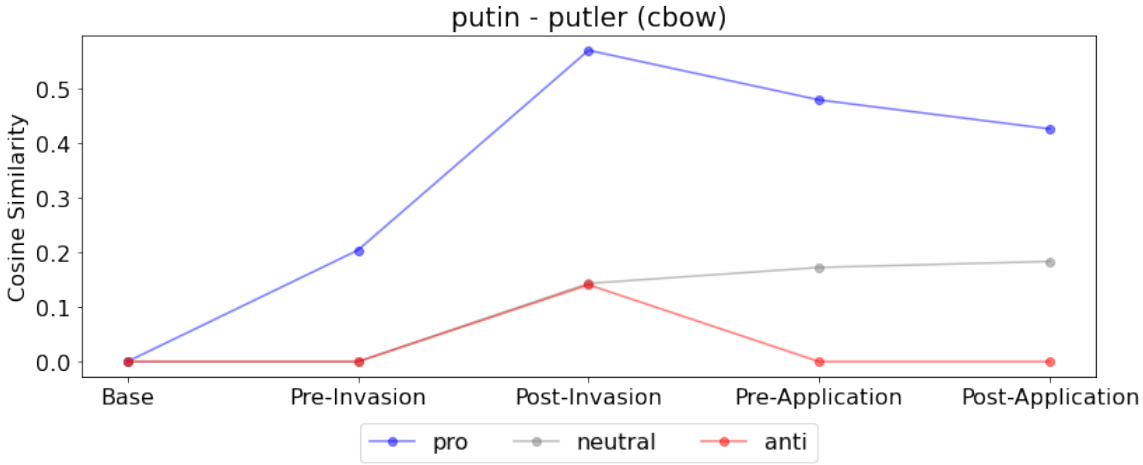

Figure 4.25: The change in cosine similarity between *putin* and *putler* over time for pro-NATO and anti-NATO users.

5

Discussion

This chapter broadly summarizes the findings and analyzes the results in the context of the echo chamber hypothesis. Evidence for and against echo chambers is presented. Discussion of the limitations of the methodology and difficulties associated with the dataset follow, including examination of individual topics of discussion and representative tweets for those topics. The discussion concludes with an analysis of whether or not the user communities formed around topics of discussion rather than shared opinion.

5.1 Overview

Conclusively identifying communities which were echo chambers was not possible with the topology of the network alone using the algorithms tested on the dataset collected. The hypothesis that it is possible to detect ideologically-similar user communities based on topology alone is not supported so far by the use of modularity-based community detection or the Infomap Algorithm on a set of tweets containing a discussion about Sweden's entry into NATO.

The groups of users returned are a broad mixture of political leanings, perhaps due to communities being overly large. Modularity-based community detection algorithms are known to suffer from a resolution limit that prevents the detection of communities below a certain size relative to the overall network. This can lead to a failure to isolate politically similar subsets of a larger user group. The Constant Potts Model is not subject to the resolution limit problem, which perhaps explains why, especially in the case of the likes-only network, it found more homogeneous communities.

Hierarchical community detection did show subcommunities of more homogeneous position on the NATO question inside of the top level communities, at least in one case almost entirely separating out the anti-NATO minority of one of the larger communities. It was observed that often unique word associations that appeared within a larger community did so only because of the use of the term by a smaller subcommunity, reaffirming that the communities detected may have been too large to be good candidates for echo chambers formed around one side of this particular issue.

The motivation for using the Infomap and Leiden Algorithms was due to their ability to return communities of high quality in a strictly mathematical sense. The original paper which put forward the two algorithms as methods for finding communities from Twitter data left the question of their fitness otherwise as an open question: "Sociologically, the use of the tested algorithms poses a general question on their applicability for current social science tasks ... How does the structure of hidden communities relate to the expected social, cultural, and/or political cleavages in the discussion?" [26]. Though it is difficult to separate the method from the topical context when drawing any conclusions about the performance of the algorithms, this study has at least given some indication that the proposed algorithms do not necessarily return communities strictly divide along lines of political cleavage and that further modification to the algorithms is needed.

The hypothesis that communities of a given political leaning use language differently is only weakly affirmed by the evidence gathered in this study, primarily due to the inability to secure a

labeling of the communities beyond that which was achieved via a very sparse hand annotation of only 2.5% of the data, leaving the true leaning of the community uncertain. General patterns were observed for communities and users which were rated as pro-NATO or anti-NATO, but whether those ratings and thus the patterns of language use would have held had the rest of the tweets been annotated is unknown. The difference in word embeddings appeared to be driven more by the topic of discussion than by small nuances in language interpretation.

It is also not certain that the echo chambers sought existed at all among the users who contributed to the tweets in the dataset. For another subject, with clearer divides, the approach used may have returned clearly disconnected communities of distinct political opinion.

5.2 Word Embeddings and Echo Chambers

5.2.1 Observations of Echo Chambers

The vector word embeddings found as a result of this work so far show a distinct difference in language use between the pre-trained corpus and the Twitter data *as a whole*, but only weak and sporadic differences between individual groups. These languages changes showed thematic shift, as words associated with the Crimean Peninsula became less strongly associated with geographically similar terms, such as the Kerch or Taman Peninsulas, and more with politically similar terms–in this case the Donbass region of Ukraine and Abkhazia region of Georgia which received a similar treatment in conflicts between the Russian Federation and its neighbors.

Between communities detected by the community detection algorithms, the differences were less clear. When looking at the most similar words, in terms of cosine difference, few noteworthy differences appeared. Terms such as *nato*, *sverige*, and *alliansfritt* had fairly ”neutral” words as their most similar embedding, giving no indication of the position of the community’s opinion of NATO. Even more loaded phrases such as *krigsallians* compared to *försvarallians* only had weak indications of community leaning.

Those differences that did appear did so because of a topic of conversation that was popular within the community but not in others, and these terms did represent somewhat clear positions on the NATO question. Discussion of the Azov Battalion’s Neo-Nazi ties or mockery of Nooshi Dadgostar’s fears of nuclear weapons near her child’s school did not occur in all communities, so close cosine similarity between terms such as *azov* and *nazister* or *forskolor* and *kärnvapen* only appeared in some communities. However, then though *nazister* was the most similar term to *azov* in several communities, the cosine similarity between the words was not particularly high, and indeed not very much higher than in those communities where it was not one of the most similar words.

Some general trends were observed despite the lack of conclusive labeling. For example, comparisons between Vladimir Putin and Adolf Hitler were stronger in communities rated as being pro-NATO, which also had a lower cosine similarity between *ingå* and *krigsallians*, illustrating a tendency for some users’ opinion of Sweden joining NATO to be based upon their opinion of Russia’s actions in Ukraine.

Such trends were only observed because communities had been manually annotated to give an indication of pro- or anti-NATO positions, and only the clearest case were compared. Even amongst these, not all indications of pro- or anti-NATO positions appeared in every community with a pro- or anti-NATO rating. As was shown, these relationships in themselves did not provide clear clustering of communities into pro- or anti-NATO positions.

Some of the methods proposed for finding word pairs whose cosine similar could potentially indicate a community’s position on the issue proved fruitful, discovering a connection between conspiracy theories around the World Economic Forum and opposition to NATO. However, most of

the word pairs returned by these methods proved uninteresting, requiring a large amount of manual investigation.

5.2.2 Quality of Tweets and Dataset Size

The use of NLP on social media data faces some unique challenges when it comes to the quality of the dataset. Twitter users do not always subscribe to correct grammar or spelling and often use conventional abbreviations. It was not possible to clean the text to normalize language in all of these cases, and many variations in spelling or abbreviation occurred too seldom to form meaningful relationships with the canonical word. Because of this, many word embeddings may not truly reflect the users' intented meaning.

> FUCK nato och erdogan och ann linde och sossarna jävla RÅTTOR

> Både Turkiet o Kroatien säger alltså nej till ett svenskt 🤭 medlemskap i Nato. 😂 😂 😂 😂 Angående Sverigebilden idag 🦍 🦍 🦍 🦍 🦍 Kan inte sluta skratta 😂 😂 Heja 🤭 🤭 sossezverige 😂 (S) 😂 😂

> M A's meriter 2022 : utb botten EU vanlig sjukv dito åldr.vård katastrof sämst pension Norden, botten EU krim värst i EU neutralitet såld ej folkomröstad besökt Ukraina som förbj rysk minor tala sitt språk USA-marionett fattar ej själv men kvinna bra sim

Additionally, many tweets contained more information in attached pictures or links to websites than they did in their own next. Users would commonly post screenshots of other text, and this study did not make any attempts to analyze image attachments. Future studies of social media data which contain multimedia attachments would benefit from finding ways to include this information. It was necessary to visit many of the included links to be able to manually annotate a tweet to fully understand the context. Nearly a quarter of tweets contained an attachment.

Breaking the full dataset of 19.9 million tokens into smaller sets, some with fewer than 1 million tokens, meant that in several of the communities words of interest appeared too few times to secure meaningful word embeddings, which required a minimum of 10 appearances of any token to be considered in the models employed here. This compounds the issue of irregular spelling, misspellings, and abbreviations.

Interesting echo chambers may have been present in the data, but perhaps only in the smallest communities which had insufficient text output to put their tweets through the Word2Vec process. A large number of these communities, ranging from dozens to a few hundred users, were omitted from further analysis after the community detection algorithms were run due to their size.

5.3 Variables Considered

The conclusions drawn from this work should be considered for only the methods used and not taken as evidence for or against the general suitability of community detection and subsequent word embedding analysis for echo chamber detection as all possible configurations were not exhaustively explored. This study incorporated several steps, from data gathering to community detection and finally the use of word embedding models, where in each step only a small number of configurations

were considered based on evidence that was available in support of the method. The impact of these selections on the results, as well as some possible modifications, are discussed in this section.

5.3.1 User Network

Two network topographies were considered. The first used all available connections, including users following other users, replying to tweets, retweeting, or quoting other users. The second only considered connections based on when one user liked another user's tweet. The latter network topography seemed to make more sense for finding ideologically similar communities, as users who liked each others tweets would be more likely to share an opinion than just those who followed or replied. However, the resulting network structure gave poorer communities (in terms of the considered metrics) as a result. Connections had a tendency to be only one-way; a small number of users generated content with a large number of likes, and most other users had comments without any likes at all. Around 1/3 of all tweets had no other users liking them, and fewer than 1/10 had more than 100 users liking them. The sparseness of the graph and worse quality partition suggests that some other weighting could have been considered. For example, if likes were weighed more heavily, while other connections still were retained as edges, then perhaps different results would have been found. As prior research has shown, not all connections are indicative of shared opinions, but users do have a tendency to follow sources that they hold in higher regard so those connections should be somehow considered. A set of guidelines for how to determine the relative weight of different types of user connections would benefit this kind of research.

5.3.2 Classification of Communities

The classification of users and communities as pro- or anti-NATO performed here is also only one of many methods that could have been used. Other researchers that investigated echo chambers in a similar approach have managed to classify communities' and users' political positions using other methods, such as finding media sources that users follow and inferring a bias score for each user from that[3], rating the polarization of media that users linked to[49], or searching for hashtags while drawing from a much larger dataset that is trimmed down to eliminate non-politically active users[50].

We should be careful not to draw any conclusions from the fact that the network assembled from only from users like other users' tweets resulted in communities with more homogeneous "NATO scores" than the full network, as the scoring method took only likes into account, and only 2% of tweets were scored. If all tweets had been scored, or even a majority, then user's scores would be more heavily weighted to their own postings and some conclusions could be drawn. Further research involving a fully annotated dataset could give useful insight into how social media networks can be constructed so that community detection algorithms are able to detect ideologically-similar communities.

5.3.3 Word Embedding Models

Considering both skip-gram and CBOW side-by-side proved fruitful. The skip-gram model highlighted correlations between word vectors that did not appear at all in the CBOW model embeddings, while the CBOW model gave more reasonable word embeddings for the more common words in the dataset. What cannot be ruled out is that additional word embedding tools, such as GLoVe or FastText, would also provide evidence of novel word uses that both Word2Vec models missed, though it is not expected that wildly different results would be found using those methods.

Lemmatization of the Twitter dataset proved to give worse results when cross-checking for known relationships. For example, the word *vilde* had very specific meaning in a political context is lemmatized to *vild*; although related, the two words do not share the same connotation. This makes the relationship with *vilde* misleading. No advantage to lemmatization was seen when looking through the data in that no new or novel word similarities were found when analyzing word embeddings of the lemmatized text. It can be assumed that for this type of task, specifically in Swedish, lemmatization is not a beneficial step in the pre-processing chain. Lemmatization was the most computationally expensive step in the pre-processing chain.

The initial inclusion and subsequent rejection of lemmatization and lack of a clearly better choice between two Word2Vec models and two network topographies are an illustration of the obstacles in this research. When attempting to develop a method to detect what is being said within echo chambers, many different variables need to be considered to rule them out–lemmatize or not, continuous bag of word or skip-gram, Word2Vec or GLoVe, or any other potential avenues of investigation. A fully exhaustive exploration of the data considering multiple word embedding models and additional community detection algorithms would require a prohibitively long amount of time to perform. It may be that a different model, or different parameters, or a different partition would find conclusively distinct uses of language in different communities.

5.4 Do Echo Chambers Exist?

5.4.1 Echo Chambers and Social Media

Several possible reasons exist for the lack of clear echo chambers in the data. One explanation is that the selective exposure hypothesis is wrong; users are not only following other users who share their opinions, so user feeds are populated with perspectives that they disagree with almost as often as those that they agree with. This conclusion would be in agreement with Törnberg's paper[18] that some other mechanism than selective exposure is responsible for political polarization. Factors such as education, income, age, sex, ethnicity, etc., may be as likely to influence a user's opinion about NATO membership as who they choose to follow on Twitter is, but personal factors were not collected for users for this study.

Supporting this is the fact that Twitter uses an algorithm to personalize the feed of tweets that users see when they log in that may expose them to information from users that they did not choose to follow. This algorithm suggests content to users, regardless of whether they are following the poster or not, based upon the popularity of the tweet and how users in one's network are engaging with it, as well as user's own posting habits[49]. This immediately casts doubt upon the selective exposure hypothesis, as Twitter may be forcing users to see content supporting positions that they disagree with in order to drive engagement. It is a pattern seen through other studies that social media users are actually being exposed to more diverse content than they would otherwise encounter [18].

In fact, it is observed that there is a considerable amount of communication between the most strong pro-NATO and ant-NATO rated users. Many of the communities detected by the algorithms used here combine pro- and anti-NATO users into one community, meaning that those users have greater communication with their ideological opponents than they do with other similarly minded users in the network. Users who were rated ”strongly pro-NATO” were three times as likely to follow other users who were rated ”strong pro-NATO” than users rated ”strongly anti-NATO”, however they were more than twice as likely to follow ”strongly anti-NATO” users than neutral users.

We cannot rule out selective exposure still being a factor in users' information intake if they instead seek out newspaper, radio, television, or other online new sources and do not invest the

effort to curate a feed of like-minded fellows on Twitter. None of the data collected for this study can speak to the phenomenon outside the narrow scope of a user's Twitter behavior.

Another possible explanation for communities without strong evidence of echo chambers is that this study considered a broad base of users which includes a large number of politically ambivalent Swedes who do not have very strong opinions about the issue. The majority of social media users are not active enough in a political conversation to form echo chambers[36], casting doubts on whether echo chambers–or any similarly ideologically-aligned group–can be found in the limited dataset collected. Removing all but the most prolific and opinionated users would reduce the dataset size to a point in which there would likely be insufficient data for any word embedding training. Indeed, some previous studies which based their conclusions in favor of the existence of echo chambers on only the most active and politically divisive core of users may not be capturing the true picture of polarization in social media networks[51]. A previous study which sought to characterize echo chambers on various social media platforms considered only Twitter users who had posted links to partisan news outlets[3], already a more selective criteria than the user collection method used in this work, leading to the elimination of the large number of politically ambivalent users who dominate a social media network.

5.4.2 Echo Chambers and NATO

That the subject of Sweden's entry into NATO was not widely enough discussed to have a sufficiently large core of very active and very divisive users that could be looked at in isolation raises further questions about the existence of echo chambers within the data collected.

The amount of pro-NATO content was much greater than the amount of anti-NATO content. As users with anti-NATO position are fewer and farther between (as evidenced by previously referenced polls showing a majority of Swedes in favor of joining, with the balance split between those against and those undecided or uninterested), their communities may be more scattered and less likely to be connected into a community of substantial enough size to form a cohesive echo chamber, or to appear in the data if they had. The groups that oppose NATO membership for Sweden are also quite diverse in their political otherwise. This leads to smaller communities with small text corpora, so terms unique to the anti-NATO position did not emerge as readily. Of those terms which were unique to one of the sides in the argument, as determined by the hand annotation, 686 were unique to pro-NATO communities and 33 were unique to anti-NATO communities.

The issue of Sweden's entry into NATO may not have been given sufficient time to generate any echo chambers. The possibility of Swedish entry into NATO was not earnestly discussed until the Russian invasion of Ukraine, and in less than three months Sweden had submitted a formal application to join the alliance. Twitter activity was greatest immediately following major events, such as the submission of the application or the vote of no confidence, but no sustained discussion existed.

Prime Minister Andersson did not engage in public debates around NATO, likely to prevent the NATO question from becoming a political one. The NATO question would be difficult politically for the Social Democrats had it become an election issue; supporters of joining NATO would have been put off by their long-held anti-NATO stance, and anti-NATO voters would opt for the Left Party or the Green Party which both still opposed joining NATO.

The timeline of Twitter activity has shown relatively low engagement with the topic after the NATO application, and no pre-election bump as would have been seen had it been an election issue. Figure 5.1 shows the increase in tweets containing crime-related terms as the election nears. Only a small, brief spike is seen in the NATO-related terms. Crime was listed as one of the most important election issues by half the electorate on the eve of the election in September[N25]; NATO

membership did not make the list.

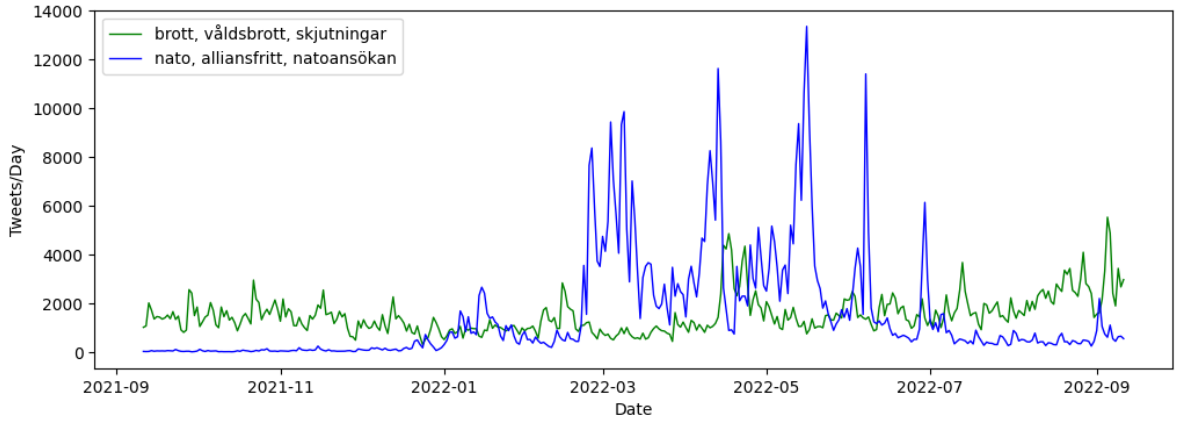

Figure 5.1: NATO-related search terms totally roughly a half million Tweets compared with crime-related search terms also reaching a half million tweets. As the September 11th election approaches, the number of tweets including crime-related terminology steadily increases.

The question of Sweden's accession to NATO alone does not seem to generate a large discussion of that topic on its own. Instead, the Twitter data collected seems to primarily focus on Sweden's NATO involvement as a factor in other contexts, most commonly the War in Ukraine, which users often cast as a war between Russia and NATO. NATO-related tweets generally refer to the potential consequences of Sweden joining or not joining NATO in terms of the ongoing conflict, and it does not seem possible to separate the two.

Some other positions that users commonly took related to the NATO question follow as an illustration of the multiple facets of the NATO discussion which may have contributed to no clear echo chambers being observed.

5.4.2.1 Urgency in Defense of Sweden

The primary argument for NATO membership was the most obvious one: protection. Users in favor of NATO membership argued that NATO would protected Sweden from being attacked by Russia. These users saw a need to join due to the threat that Russia posed, commonly implying that Sweden was set to suffer the same fate as Ukraine, while many users saw no such attack on the horizon and felt membership was unnecessary. For many of those, it was believed that EU membership or informal partnerships with the alliance would be sufficient for Sweden's protection.

Here, the urgency of Sweden's NATO bid is pushed by the pro-NATO side. Closer relationships seen in the word embeddings between *putin* and *hitler*, and between *frihet* and *säkerhet* suggests that the pro-NATO users see Sweden as being actively under threat. Here also the stronger relationship between *skydd* and *försvarsgarantier* as well as *ensam* and *stark* indicates that the pro-NATO side not only sees a threat, but feels Swedish is incapable of handling the threat on its own.

> Det kanske enskilt viktigaste lärdomen av vad som nu sker i Ukraina, är den avgrundsdjupa skillnaden mellan att vara medlem i Nato eller enbart partner. Krävs inte mycket för att förstå varför det är Ukraina och inte de baltiska staterna som Putins bomber nu faller över.

> Vi behöver inte vara med i den politiserade och svindyra organisationen NATO. Vi samarbetar redan med dem utan alla nackdelar att vara medlem i NATO.

Many users appealed to neutrality as the default and safe option for Sweden, alluding to a belief that neutrality had kept Sweden out of both World Wars and the Cold War conflict. Some users commented as to how the current conflict in Europe represents less of a threat of war than the Cold War did. Common pro-NATO arguments were that Sweden was not truly neutral, and that neutrality was no guarantee of protection. The difference here is underscored by the lack of use of the term *allianslös* and its inflections by anti-NATO users and communities.

> Vår neutralitet har gett oss fred i över 200 år.

> Förklara på vilket sätt vår ”neutralitet” skulle stoppa Putin om han vill använda Sverige som spelbricka. ETT vettigt argument..?

> Sverige måste gå med i Nato genast, för vår egen skull. Att vara allianslös är inget skydd. Du vet hur Ukraina har det, eller hur?

5.4.2.2 Destabilization, Ukraine, and NATO Expansion

A further area of debate was whether NATO's expansion is a provocation directed towards Russia, one that was answered by the invasion of Ukraine in order to maintain a buffer state. Many users believed that war in Fennoscandia would be best avoided by not antagonising Russia by applying for NATO membership. These arguments were countered by users who believed that allowing threats from Russia to dictate our policy was unacceptable. Not all tweets on this topic applied to Swedish membership directly; many discussed the war between Russia and Ukraine in terms of NATO expansion and encirclement.

> Självklart är Natos expansion precis det som Ryssland upplever som hotfullt och det som de reagerar på.

> De som argumenterar för att NATO-medlemskap skulle trigga Putin, gör precis det Putin vill. Skrämma oss. Är vi inte med i NATO får vi inte militärt skydd. Så enkelt är det. Det finns inget bättre läge än nu att skicka en skarp signal till Moskva att deras hot inte biter på oss!

Particularly interesting here is the connections between *2014* and either *annekteringen* on the pro-NATO side or *statskuppen* on the anti-NATO side. Users who were in favor of joining NATO displayed a tendency to cast Russia as an unprovoked aggressor invading a peaceful and diplomatic Ukraine, while many of those on the anti-NATO side saw the unrest in Ukraine in 2014 as part of a wider struggle by the West to pull the nation out of Russia's sphere of influence. The relationships between *azov* and *nazister* also suggests a willingness to consider Russia's point of view in the matter and shows users who are less likely to see Ukraine as an innocent victim of unprovoked aggression. If one believes that there is a justification for Russia's invasion of Ukraine other than

mere aggression, then the argument that Sweden is also in danger simply by being close to Russia is weaker.

> NATO hade mening under kalla kriget. Nu är det inte en försvarallians och när Ryssland blev självständigt med Putin har NATO använts av USA för att omringa och försöka knäcka Ryssland. Allt för att få bort Ryssland som självständig makt.

> @GostaHulten NATO är en kriminell organisation med blod av många nationer på sin hand. Ukraina hade ca 5000 soldater, vara en del ny nazister från Azov Battalion, i Irak. Att Sverige går i NATO betyder att Sveriges öde blir som UA-öde. Våra liv är i fara om Sverige gå med i NATO.

In light of this, the question to join NATO, once it had been proposed, was for many a choice between NATO and Russia with no middle ground. Joining NATO was seen as siding with Ukraine against Russia, while opposition to NATO meant not just neutrality, but support for Russia's actions in many users' eyes.

> Finns inget annat alternativ än att rösta bort den här Putinvänliga regeringen vars kvinnor har sagt nej till Nato och röstat för Putin.

> Det är ju snarare så att det är en stalinkramande diktator som invaderar en självständig demokratisk stat. Fortsätt du att propagera för neutralitet men du lär nog får svårt att övertyga någon större skara.

5.4.2.3 Obligations, Cost of NATO, and Turkey

Man users were engaged in a debate not about whether joining NATO was necessary, but if it was worth the costs, both financially and morally.

Some users entertained the possibility of Sweden handling its own defense, but saw NATO as a more cost-effective way to provide for defense. Frequent references to the guideline that all member nations should spend 2% of their GDP on defense appeared alongside the 3-4% figure. The Word2Vec models had found the relationship between *2%* and *bnp*, but there was no clear divide regarding an association with either a 3-4% figure or other spending estimates between the two sides.

> Är det negativt att vi skulle spara pengar på försvaret med att gå med i NATO? Det skulle kosta betydligt mer än 2% att ha ett eget försvar på helt egna ben. Jämför med utgifter under 50-70 tal med 3-4% / år.

Another more opinionated point of opposition was concern for Turkey's Human Right's record regarding the Kurds. Users were quick to draw comparisons between Turkey's foray into Syria and Russia's invasion of Ukraine, and saw joining an alliance with Turkey as a problem. The similarity between *erdogan* and *diktatur* was observed across the entire dataset.

> Erdogan och Putin bedyrar sin vänskap, var inte ryssen anledningen till att vi skulle gå med i Nato? Sedan är ju Turkiets behandling av kurder och oliktänkande en anledning att bojkotta Turkiet, inte joina samma klubb…

The argument here was about whether Turkish demands for Sweden to join NATO were a price worth paying. Many users believed it was not worth sacrificing the principles of democracy and freedom of expression, which many users felt would need to be compromised in order to satisfy Turkey's demands.

> För mig är detta en fråga om att behålla Sveriges autonomi. Hela anledningen att vi ska gå med i NATO är att skydda oss mot potentiell påverkan från ett ofritt Ryssland.
> Om priset för det är att acceptera faktisk påverkan från ett ofritt Turkiet är det för högt.

Others took a more pragmatic position in their opposition to joining an alliance with Turkey. These users saw joining NATO as a way to force young Swedes to die for Turkey's sake.

> Det är inte värt att dö för Turkiet och Erdogan. Natomedlemskap är därför inte rätt väg.

These users might have agreed that Sweden needed NATO protection, or that NATO expansion was not a threat to Russia, however they end up on the anti-NATO side of the argument for other positions. This makes it difficult to see any collection of anti-NATO users as a monolith.

For many users, it was the US which served as the main obstacle for joining NATO. Opposition ranged from the presumption that the US was somehow coercing Swedish leadership into a NATO application, or the joining NATO would lead to dependency on the US and partial responsibility for US war crimes. That some of the most strongly anti-NATO communities also had the highest cosine similarity between *nato* and *usa* is evidence towards this.

> 40 år sen Spanien gick med i #Nato Dåvarande stabminister José Manuel Otero Novas avslöjar nu att USA hotade med att göra Kanarieöarna självständiga om #Spanien vägrade gå med i Nato

> Ska Sverige sälja ut mänskliga rättigheter för att bli medlem i USA:s krigsorganisation (Afghanistan, Libyen,Irak)?

5.4.2.4 Political Infighting

A characteristic position conveyed in tweets with apparent pro-NATO leaning was to cast Prime Minster Magdalena Andersson as irresponsible by accepting Amineh Kakabaveh's support in Morgan Johansson's no-confidence vote in exchange for public statements of support for Kurdish groups. The tweets complained that Andersson was jeopardizing or sabotaging the NATO application by this move, implying that the users were in favor of the NATO application. However, the main position here is not support for Sweden's entry into NATO, but rather casting doubt on the leadership abilities of the Social Democrats leading up to the election.

> Så Magdalena Andersson valde till slut att äventyra landets Natoansökan för att blidka Amineh Kakabaveh - och därigenom rädda kvar Morgan Johansson. Ett totalt ansvarslöst agerande av en statsminister som aldrig mer borde ta ordet ansvar i sin mun.

Common themes in these types of posts were to cast the Social Democrats as prioritizing internal party issues over global issues, meant to cast doubt on the responsibility of the party to do what was best for Sweden as a whole. This triggered responses from users who saw the Vote of No Confidence as an irresponsible political maneuver.

> Det är inte direkt snyggt, och det är definitivt inte till Sveriges fördel när Magdalena Andersson medvetet försöker skapa en känsla av politiskt kaos. ”Hon gör partipolitiskt spel av Nato-ansökan”

> Ulf Kristersson anklagar Magdalena Andersson för politiskt spel och att inte vilja fokusera på sakpolitik - efter att hans parti vill kasta i Sverige i en regeringskris mitt under pågående Nato-ansökan, tre månader innan ett val.

Other types of tweets related to the NATO issue were those which took specific aim at Swedish Defense Minister Peter Hultqvist, who at one point had promised that Sweden would never join NATO while he was in the government. This about-face attracted a lot of mockery on Twitter, though it is impossible to tell if the tweets were made by those with anti-NATO positions who were disappointed, or those with pro-NATO position who were looking for an reason to attack the Social Democrats.

> Du menar att man ska lita på Hultqvist ”JAG KOMMER ALDRIG GÅ MED I NATO” och Löfven ”MITT EUROPA BYGGER INGA MURAR”?

It was not only the Social Democrats who suffered for a reversal on their NATO position. Many users posted tweets indicating that they had abandoned the Sweden Democrats party for opting for NATO membership. A common theme among posters who indicated support for AfS, an Extreme-Right fringe party in Sweden, was that the Sweden Democrats were traitors. This tweet, posted as a response to Björn Söder, a member of the Sweden Democrats party, illustrates the trend:

> Lek inte duktiga nyttiga idioter som vill gå med i Nato helt plötsligt. Ingen vill rösta på er, förbaskade Landsförrädare.

Examples of Far-Right groups which opposed Swedish NATO membership were the neo-Nazi group Nordiska Motstånds Rörelse (NMR). That the ”anti-NATO” community should include both far-left and far-right elements is perhaps one reason that clear echo chambers were not observed; these users coincidentally agree on just one issue and are opposed on so many others that their social media interactions are not expected to be frequent. This may also be a reason why the community detection algorithms found pro- and anti-NATO users within each of the major communities who were not strongly connected to users in other communities who had the same ideological position on the issue. For example, some users used the subject of Turkish opposition to make further criticisms

of the Social Democrats policies and presented Turkey's opposition as evidence that Sweden had failed to address terrorism and crime in a suitable manner.

> President Erdogan vill inte ha med Sverige i Nato pga att Sverige är ett tillhåll för terrorister, våldtäktsmän och kriminella, vilket fint land S har skapat.

A voter for the Moderate Party who is pro-NATO could as easily have made this tweet as a voter for AfS, who is most likely to be anti-NATO.

This issue of rising crime in Sweden emerged as a tangential issue in the dataset. Justice Minister Morgan Johansson was subjected to vote of no confidence in June of 2022. The no confidence motion came as a result of rising crime in Sweden, and that opposition view was that the government had taken insufficient steps to limit it. Among the twitter users who spoke on the issue of crime, a major incident that was often referenced was a series of riots which followed the buring of a Koran by Rasmus Paluden[N26]. Dozens of police officers were injured in the riots by stones being thrown, vehicles being burned, and other violent altercations with the predominantly Muslim protesters. The incident highlighted an underlying conflict between Freedom of Religion and Freedom of Expression[N27], which was a factor brought up by many users in the Twitter dataset.

An example of a tweet related to this incident is found as a parent tweet to a tweet which referenced the effect that Amineh Kakabaveh's support for Morgan Johansson would ultimately have on Sweden's NATO application:

> Upploppen igår beror inte på ”kränkta muslimer”. De är resultatet av årtionden av dysfunktionell och naiv invandrings- och kriminalpolitik.

5.4.2.5 Democracy

Other users took the position that the decision was insufficiently democratic and was being made based upon fear rather than a sound assessment of the long-term ramifications of joining NATO. Users who insisted upon a referendum generally implied that they wanted the opportunity to vote against NATO membership. Most NATO members did not hold public referenda for membership; Spain, Hungary, Slovenia, and North Macedonia are the exceptions.

> Både SD och S gick till val 2018 på nej till Nato. Sedan svängde de på några veckor. Ska man svänga så, så borde det åtminstone vara ett val emellan så att väljarna får säga sitt. Helst en folkomröstning också. Såhär stora beslut ska fattas av folket!

Still, some users pointed to the polls which showed a majority of Swedes in favor of joining NATO as all the democratic justification that an application to the alliance needed.

> Jo fast nu var en klar majoritet av Svenska folket för ett NATO-medlemskap, varför man då (och först när det var etablerat) ansökte om det. Det kallas demokrati, passar det inte helt plötsligt eftersom ”nej till NATO” -sidan torskade nu?

5.5 Topic Modeling

If the communities detected are not echo chambers or even of a consensus opinion on NATO in most cases, then perhaps they could have some other significance. As there are numerous subtopics for the NATO question, it is worthwhile to investigate whether the communities detected are not users of shared opinions, but rather users discussing the same topic. One of the smaller communities was found to be entirely composed of users discussing the ongoing conflict in Ethiopia; it possibility that other communities were also discussion groups around a single topic or small set of topics.

The method used to determine this is Structural Topic Modeling. Structural Topic Modeling is an unsupervised classification of words in a set of documents that finds natural clustering of words into a specified number of topics. One of the most common methods is to use a model based upon Latent Dirichlet Allocation which results in a probabilistic model where each document, which in this case is a single tweet, is a mixture of latent topics and each topic is a probabilistic representation of words[52].

Because LDA Topic Modeling was not initially imagined as a component of this study, the description here is kept minimal. It is sufficient to say that, for a pre-defined number of topics, each topic is expressed as a collection of the words most closely associated with the topic, and that each document is represented as a mixture of topics. A single tweet could thus have a topic placement defined as 10% Topic 1, 20% Topic 2, and 70 % Topic 3 in a 3-topic model.

The number of topics to consider is not known in advance. A general method of finding the proper number of topics is to consider models with a range of topics and look for the best results. The quality of the results is defined by two terms, the perplexity and the semantic coherence. The perplexity is a measure of how ”perplexed” a model is given a sequence of terms; that is to say, how well the model can predict the next token in a series.[53]. A lower perplexity indicates that a model is well adapted to fitting the dataset. Semantic coherence is a score which results in topics that are more readily understood by a human reader. The semantic coherence is a measure of how often top words in a topic appear together [54]. The model and topic number was selected based upon maximizing both of these scores on a corpus consisting of the entire Twitter dataset.

Structural Topic Modeling extends the LDA method by adding additional meta-data to the topic determination process, such as a time stamp for a document, in order to improve the topic grouping. It has been shown in the timeline of tweets in the dataset that there are a large number of tweets at specific points in time as a reaction to a particular event which tend to focus on the same topic. For example, tweets around February 24th are likely to be related to the Russian invasion of Ukraine, and those near May 17th are likely to be related to the NATO application being submitted. This additional data can help arrive at a collection of topics which best encompass the various issues in the developing context. The tool used for the Structural Topic Model was the STM package for R [55].

The STM derived for the Twitter dataset showed best results when the number of topics selected was 8. Of those 8 topics, 6 are clear topics based on their top words, and the remaining 2 topics were unclear. Table 5.1 shows the words with the strongest association with each of the topics and the approximate proportion of each topic within the text within the text.

Figure 5.2 shows the distribution of topic prevalence over time for topics related to party politics (Topic 1), Turkey and President Erdogan (Topic 2), and Russia and Ukraine (Topic 5). The prevalence of topics related to party politics peaks during the period of time around the Vote of No Confidence and after dropping gradually increases again towards the election. Topics related to Turkey become most prevalent after Erdogan’s first indication that Turkey was opposed to Sweden’s NATO application. Topics related to Russia and Ukraine are most prevelant in the build up and immediate aftermath of the invasion. The prevalence of these topics follows expectations based

Topic	Top Words	Proportion
1	s, fråga, nej, regering, parti	0.14
2	turkiet, erdogan, gälla, stöd, artikel	0.09
3	krig, värld, människa, sluta, hel	0.11
4	nato, sverige, svensk, finland, natomedlemskap	0.15
5	ryssland, ukraina, putin, nato, usa	0.12
6	ja, väl, fel, mången, ändå	0.15
7	måste, rätt, folk, försoka, sak	0.11
8	land, behöva, försvar, militär, stor	0.13

Table 5.1: The top words for each of the eight topics and the proportion of the text which was determined to fall under each topic.

on the dates of relevant developments.

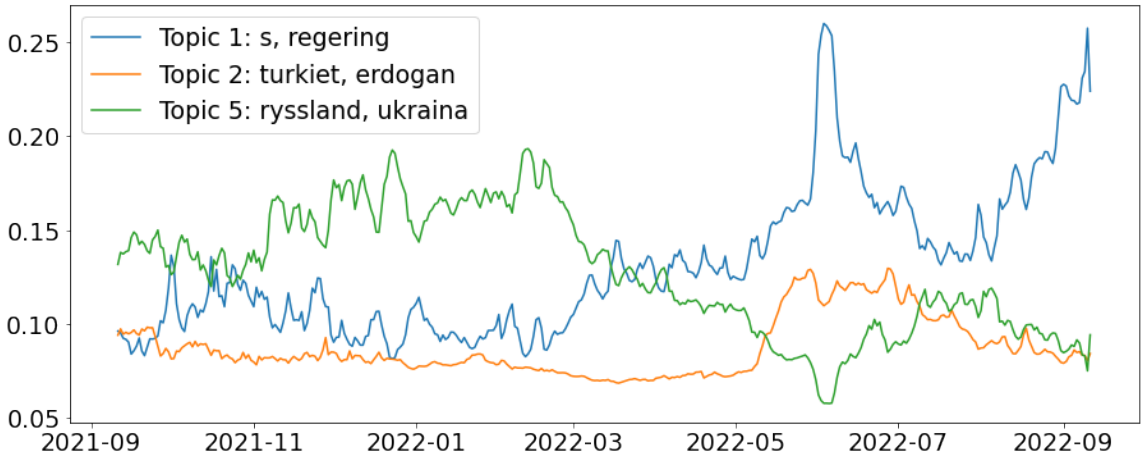

Figure 5.2: Topic distribution over time of a subset of the topics.

As an example of a tweet with a clear topical assignment, the following tweet was given a 0.73 probability belonging to Topic 1 and a low probability of belonging to any one of the other seven topics:

> Ulf: Vi samarbetade med S för NATO-ansökan Annie: Ulf sa att han kan tänka sig samarbeta med S i många frågor #valet2022

As this topic appears to be about the relationship between political parties, and Topic 1 clearly refers to party politics, the topic estimate for this tweet seems correct.

Figures 5.3 - 5.6 give the topic distribution in the communities detected by the algorithm. The proportions in the pie chart are based on the topics for which each tweet posted by a user in the community is most likely to belong to. The communities do focus on different topics, however there is not a stark difference in topic coverage except for in the smallest communities. It is clear that the communities themselves are not formed around specific topics, at least not as topics could be defined by the Structural Topic Model.

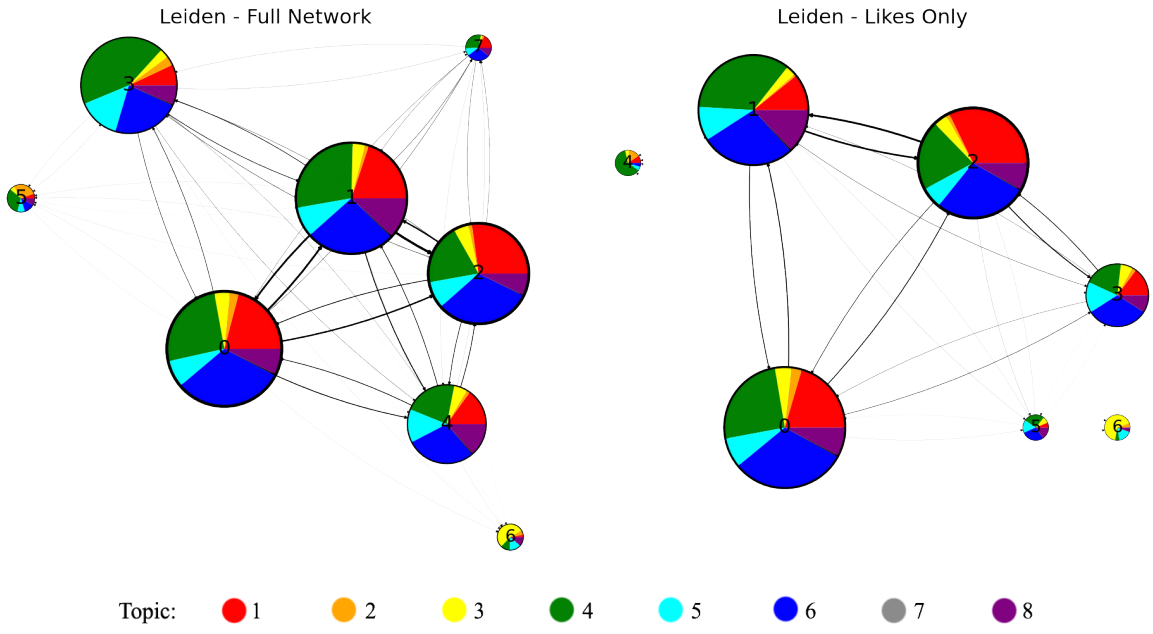

Figure 5.3: Topic distribution within the communities detected by the Leiden Algorithm.

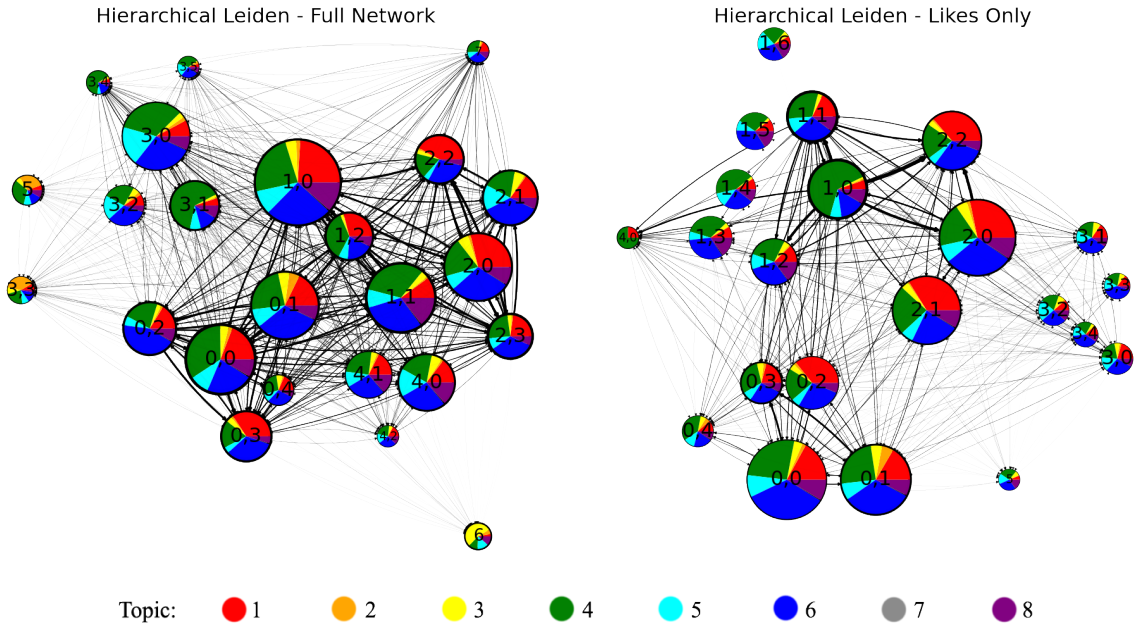

Figure 5.4: Topic distribution within communities detected by the Hierarchical Leiden Algorithm.

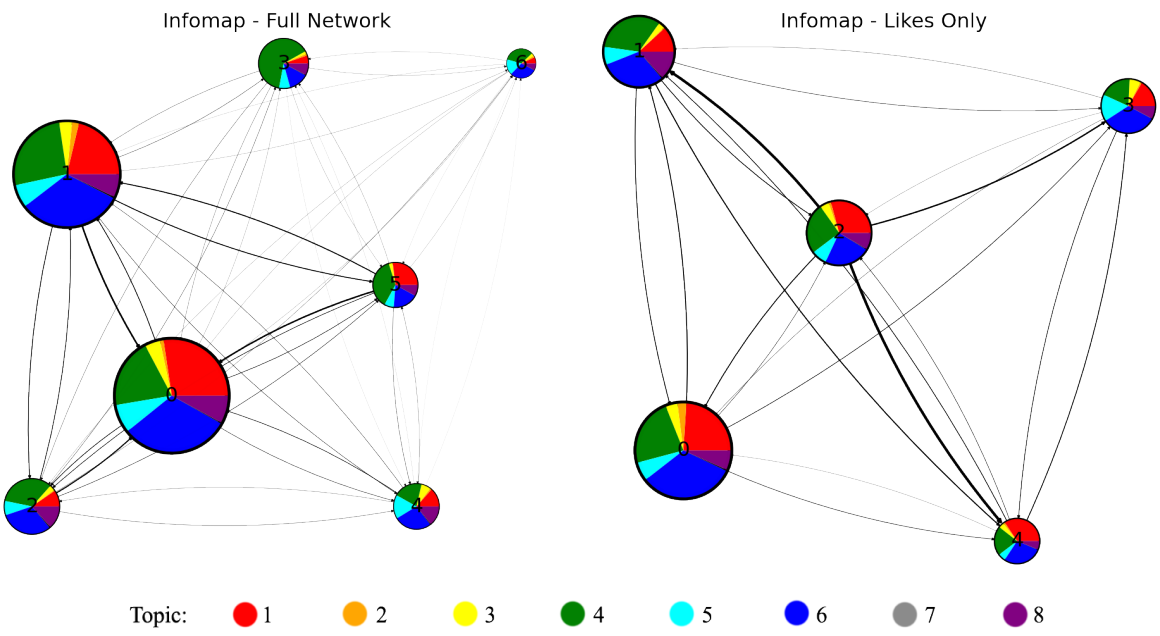

Figure 5.5: Topic distribution within communities detected by the Infomap Algorithm.

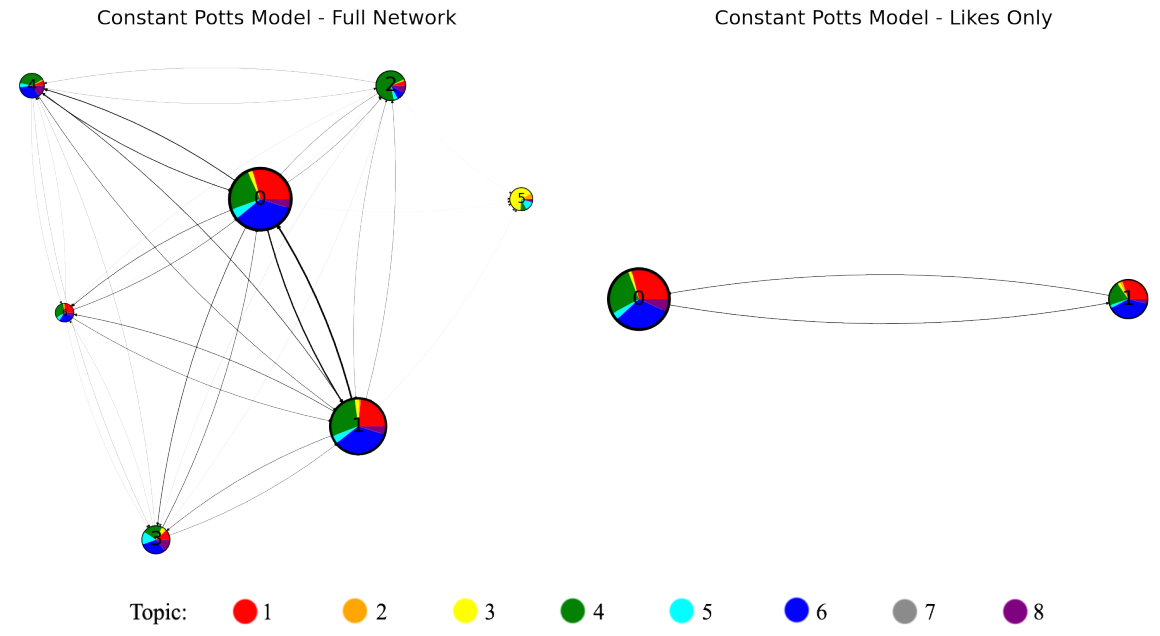

Figure 5.6: Topic distribution within communities detected by the Constant Potts Model.

6

Conclusion and Future Work

The methods proposed were able to detect some patterns of language use and hint at an association between positions on Sweden's application to NATO and the cosine similarity between some key terms, but the evidence is not conclusive and in many ways supported by an incomplete hand annotation of the dataset. It is not possible to conclude if this is a result of the choice of topic or the methods.

The three proposed hypotheses were:

1) Community groupings based on the topology of a social media network represent groups of users who will have greater similarity of opinion amongst themselves than they will with members of other user groups.

2) Language use within the community will change as a result of the differences of opinion, showing a distinction in use patterns between users who are members of one group and users who are members of another.

3) Community groupings based on the topology of a social network show a distinction in patterns of language use.

The first hypothesis is not supported by the methods considered in this study. The communities detected using both the modularity-based algorthims and the Infomap Algorithm contained several subcommunities which, when analyzed in a hierarchical manner, were quite opposed to each other on the subject of NATO despite having been included in the same large community. This is attributed to the resolution limit. The dataset itself was a problem in this regard, as breaking the network into smaller communities which were more likely to be homogeneous in ideology ended up with communities with too little text data to properly apply the Word2Vec model. What can be concluded is that, at least on this subject, those users who are most in favor of and most opposed to Sweden joining NATO are not completely isolated from each other in social media, as the echo chamber theory would suggest.

Assessment of the topical contents of each communities did not show any evidence that the communities were formed around certain subtopics of the discussion rather than ideological positions.

Evidence of the second hypothesis already exists in other research, though not on the subject of Sweden's entry into NATO. The methods proposed here have been able to detect language use patterns that correlate to a either an anti-NATO or pro-NATO position, however the communities from which those patterns were extracted could only be identified through the use of annotation of the data. Still, certain topics are discussed primarily by users who are in favor of joining NATO, while other topics are only brought up by those opposed. The word embeddings give indication of this, and the methods used were able to find some of these patterns using the results of the Word2Vec models.

Cosine differences between words of interest did not vary enough between communities detected by algorithm to offer much support for the third hypothesis. Correlations were seen in cosine similarities between selected pairs of words, however communities which showed similar values of cosine similarity were often farther removed topologically from each other than communities with very different values for cosine similarity for the word pairs in questions.

Despite not being the focus of this work or the intent of the method proposed, some interesting changes in semantic use of terms over time were observed. The topic of the study was proposed when it appeared that the NATO question would be a major election issue, and though has been seen that it was not, the time-based changes in word embeddings give some understanding about the evolution of the debate. Sweden's NATO application remains unsettled, so it may be worthwhile to revisit the topic in the future using a similar method to understand how the discussion further develops.

Future work should separate the question of whether or not the community detection algorithms proposed are able to return meaningful separations of the user network into politically or ideologically distinct communities from the problem of language analysis. For verification of the first hypothesis, a network of users with known position on an issue should be used as an input dataset for the community detection algorithms. The fitness of various algorithms in detecting ideologically similar communities can then be assessed with fewer unknowns in order to determine if such communities can be found in a social media network through this approach.

Given the dataset that has already been collected, a more complete annotation could be done to see if the findings seen so far hold true when a clearer picture of user positions on NATO is found.

A remaining question is whether or not performing the same analysis on a subset of this data, considering only those users who satisfy some criteria for engagement or strength of opinion on this topic, would yield stronger evidence of echo chambers that could be found through the graph topology even if the resulting communities were too small to extract meaningful word embeddings with Word2Vec. Several other studies on echo chambers have focused on the most divisive and active users and had found stronger evidence of echo chambers in their data as a result.

Putting aside the echo chamber discussion, the degree to which users are influenced by the language use of those whose content they follow, like, or retweet is another question which could be explored with the dataset collected. This study focused on echo chambers, considering only one way in which that influence can be seen. Tracking the emergence and spread of a single novel use of a word through the user network over time could also yield interesting insights into the flow of language influence. After all, a refutation of the echo chamber hypothesis does not mean that one's social media diet has no influence on their opinions, or that an individual's opinions are not reflected, consciously or unconsciously, in their use of language.

News Media

[N1] W. N. Glucroft, "NATO: Why Russia has a problem with its eastward expansion," *Deutsche Welle*, Feb. 2022. [Online]. Available: https://www.dw.com/en/nato-why-russia-has-a-problem-with-its-eastward-expansion/a-60891681

[N2] C. Bildt, "Are sweden and finland moving to apply for nato membership?" *Washington Post*, Mar. 2022. [Online]. Available: https://www.washingtonpost.com/opinions/2022/03/16/are-sweden-finland-moving-apply-nato-membership/

[N3] L. Roden, "What exactly is Sweden doing in North Korea?" *The Local*, Aug. 2017. [Online]. Available: https://www.thelocal.se/20170814/what-exactly-is-sweden-doing-in-north-korea/

[N4] "More swedes show support for nato," *The Local*, Jan. 2015. [Online]. Available: https://www.thelocal.se/20150109/more-swedes-show-support-for-nato/

[N5] V. Chadwick, "Swedish-Russian relations enter deep freeze," *Politico*, Sep. 2015. [Online]. Available: https://www.politico.eu/article/swedish-russian-news-nato-wallstrom/

[N6] W. Booth, "Ukraine's parliament votes to oust president; former prime minister is freed from prison," *The Washington Post*, Feb. 2014. [Online]. Available: https://www.washingtonpost.com/world/europe/ukraines-yanukovych-missing-as-protesters-take-control-of-presidential-residence-in-kiev/2014/02/22/802f7c6c-9bd2-11e3-ad71-e03637a299c0_story.html

[N7] "Ukraine protests after Yanukovych EU deal rejection," *BBC*, Dec. 2013. [Online]. Available: https://www.bbc.com/news/world-europe-25162563

[N8] O. Grytsenko, "Yanukovych confirms refusal to sign deal with EU," *Kiev Post*, Nov 2013.

[N9] N. R. Smith, "Demystifying yanukovych's decision to not sign the association agreement," *New Eastern Europe*, Nov 2013.

[N10] D. M. Herszenhorn and A. E. Kramer, "Russia offers cash infusion for Ukraine," *New York Times*, Dec 2013.

[N11] J. Henley, "Finland and Sweden take major step towards joining NATO," *The Guardian*, Apr. 2022. [Online]. Available: https://www.theguardian.com/world/2022/apr/13/finland-and-sweden-could-apply-for-nato-membership-in-weeks

[N12] "Ukraine to seek NATO membership, says PM Yatsenyuk," *BBC*, Aug 2014. [Online]. Available: https://www.bbc.com/news/world-europe-28978699

[N13] M. Champion, "Why the Minsk Accords failed to bring Ukraine peace," *The Washington Post*, Feb. 2022. [Online]. Available: https://www.washingtonpost.com/business/why-the-minsk-accords-failed-to-bring-ukraine-peace/2022/02/22/dce921da-93f7-11ec-bb31-74fc06c0a3a5_story.html

[N14] E. Yousuf, "Därför avgår inte Peter Hultqvist – trots Nato-löftet," *Göteborgs-Posten*, May 2022.

[N15] G. Tétrault-Farber and T. Balmforth, "Russia demands NATO roll back from East Europe and stay out of Ukraine," *Reuters*, Dec 2021. [Online]. Available: https://www.reuters.com/world/russia-unveils-security-guarantees-says-western-response-not-encouraging-2021-12-17/

[N16] L. Carlén and J. Olsson, “Natomedlemskap skulle destabilisera säkerhetsläget,” *SVT Nyheter*, Mar. 2022.

[N17] S. Jacobsen and J. Ahlander, “Russian invasion of ukraine forces swedes to rethink nato membership,” *Reuters*, Mar. 2022. [Online]. Available: https://www.reuters.com/business/media-telecom/majority-swedes-favour-joining-nato-poll-2022-03-04/

[N18] J. Tanner and S. Fraser, “Turkey’s leader opposes letting Finland, Sweden join NATO,” *Associated Press*, May 2022. [Online]. Available: https://apnews.com/article/russia-ukraine-middle-east-turkey-moscow-sweden-49d5297a0dff391e5de9f24f6b3a390a

[N19] A. Ringstrom and S. Johnson, “Swedish government survives no-confidence vote with help of former Kurdish fighter,” *Reuters*, June 2022. [Online]. Available: https://www.reuters.com/world/europe/swedish-mp-kakabaveh-says-she-wont-vote-against-justice-minister-2022-06-07/

[N20] “Sweden, Finland must send up to 130 ”terrorists” to Turkey for NATO bid, Erdogan says,” *Reuters*, Jan. 2023. [Online]. Available: https://www.reuters.com/world/sweden-finland-must-send-up-130-terrorists-turkey-nato-bid-2023-01-16/

[N21] A. Vatanda, “Another blow for Turkey’s free press as leading journalist arrested,” *The Independent*, Oct. 2015.

[N22] A. Kucukgocmen, “Sweden’s blocking of Turkish man’s extradition ”very negative”, ankara says,” *Reuters*, Dec 2022.

[N23] D. E. Sanger, “Biden predicts Putin will order Ukraine invasion, but ‘will regret having done it’,” *New York Times*, Jan. 2022. [Online]. Available: https://www.nytimes.com/2022/01/19/us/politics/biden-putin-russia-ukraine.html

[N24] “Dadgostars svar – efter kritiserade kärnvapenuttalandet,” *Expressen*, Apr. 2022. [Online]. Available: https://www.expressen.se/tv/nyheter/dadgostars-svar-efter-kritiserade-karnvapenuttalandet/

[N25] “Väljarnas viktigaste frågor,” *SVT*, Sep. 2022. [Online]. Available: https://www.svt.se/datajournalistik/valu2022/valjarnas-viktigaste-fragor/

[N26] A. Cassidy, M. Salem, C. Faraj, O. Nafaa, and J. Bantock, “Dozens injured in riots in Sweden after Quran burnings,” *CNN International*, Apr. 2022.

[N27] M. Nilsson, “Justitierådet: Koranbränning behöver inte innebära hat mot folkgrupp,” *SVT Nyheter*, Apr. 2022.

Bibliography

[1] J. Fernquist, L. Kaati, R. Schroeder, N. Akrami, and K. Cohen, *Twitter Bots and the Swedish Election*. Springer International Publishing, Aug. 2020, pp. 141–163. [Online]. Available: https://doi.org/10.1007/978-3-030-41251-7_6

[2] T. Fujiwara, K. Müller, and C. Schwarz, "The effect of social media on elections: Evidence from the united states," National Bureau of Economic Research, Tech. Rep., May 2021.

[3] M. Cinelli, G. D. F. Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini, "The echo chamber effect on social media," *Proceedings of the National Academy of Sciences*, vol. 118, no. 9, Jan. 2021. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2023301118

[4] G. Weimann and A. B. Am, "Digital dog whistles: The new online language of extremism," *International Journal of Security Studies*, vol. 2, 2020. [Online]. Available: https://digitalcommons.northgeorgia.edu/ijoss/vol2/iss1/4

[5] What is nato? [Accessed Oct. 23, 2022]. [Online]. Available: https://www.nato.int/nato-welcome/index.html

[6] SIPRI military expenditure database. [Accessed Oct. 11, 2022]. [Online]. Available: https://milex.sipri.org/sipri

[7] C.-G. Scott, *Swedish Social Democracy and the Vietnam War*. Elanders, 2017.

[8] J. Trudeau, "Statement by the Prime Minister of Canada on Pastor Lim's release," Office of the Prime Minister of Canada, Aug. 2017. [Online]. Available: https://pm.gc.ca/en/news/statements/2017/08/10/statement-prime-minister-canada-pastor-lims-release

[9] "Treaty of Lisbon amending the Treaty on European Union and the Treaty establishing the European Community. Article 42(7)," Dec. 2007.

[10] V. M. Villa G, Pasi G, "Echo chamber detection and analysis: A topology- and content-based approach in the COVID-19 scenario," *Soc Netw Anal Min.*, vol. 11, no. 1, 2021.

[11] A. Anagnostopoulos, A. Bessi, G. Caldarelli, M. D. Vicario, F. Petroni, A. Scala, F. Zollo, and W. Quattrociocchi, "Viral misinformation: The role of homophily and polarization," *CoRR*, vol. abs/1411.2893, 2014. [Online]. Available: http://arxiv.org/abs/1411.2893

[12] C. Sunstein, "The law of group polarization," *Journal of Political Philosophy*, vol. 10, pp. 175 – 195, Dec. 2002.

[13] A. A. Anderson, D. Brossard, D. A. Scheufele, M. A. Xenos, and P. Ladwig, "The "Nasty Effect:" Online Incivility and Risk Perceptions of Emerging Technologies*," *Journal of Computer-Mediated Communication*, vol. 19, no. 3, pp. 373–387, Apr. 2014. [Online]. Available: https://doi.org/10.1111/jcc4.12009

[14] H. Aujla, "Language experience predicts semantic priming of lexical decision," *Canadian Journal of Experimenal Psychology / Revue canadienne de psychologie expérimentale*, vol. 75, 2021.

[15] J. Newman, *The Routledge Handbook of Semantics*, N. Riemer, Ed. Routledge, 2016.

[16] C. A. Bail, L. P. Argyle, T. W. Brown, and A. Volfovsky, "Exposure to opposing views on social media can increase political polarization," *Proceedings of the National*

Academy of Sciences, vol. 115, pp. 9216–9221, 2018. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1804840115

[17] R. K. Garrett, S. Gvirsman, B. Johnson, Y. Tsfati, R. Neo, and A. Dal, "Implications of pro- and counterattitudinal information exposure for affective polarization," *Human Communication Research*, vol. 40, Apr. 2014.

[18] P. Törnberg, "How digital media drive affective polarization through partisan sorting," *Proceedings of the National Academy of Sciences*, vol. 119, no. 42, p. e2207159119, 2022. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2207159119

[19] S. Simpson, N. Adams, C. Brugman, and T. J. Conners, "Detecting novel and emerging drug terms using natural language processing: A social media corpus study," *JMIR public health and surveillance*, vol. 4, no. 1, Jan. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5838358/

[20] B. Noble, A. Sayeed, R. Fernández, and S. Larsson, "Semantic shift in social networks," in *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, Aug. 2021, pp. 26–37. [Online]. Available: https://aclanthology.org/2021.starsem-1.3

[21] A. Fredén, M. Johansson, P. K. Merino, and D. Saynova, "A comparison of language processing models in political analysis: Evidence from Sweden," ser. 2021 APSA Annual Meeting, Oct. 2021.

[22] T. Mikolov, K. Chen, G. Corrado, J. Dean, and I. Sutskever, "Distributed representations of words and phrases and their compositionality," Oct 2013. [Online]. Available: https://arxiv.org/abs/1310.4546

[23] P. Fallgren, J. Segeblad, and M. Kuhlmann, "Towards a standard dataset of Swedish word vectors," in *Proceedings of the Sixth Swedish Language Technology Conference (SLTC)*, Umeå, Sweden, 2016.

[24] G. C. Tomas Mikolov, Kai Chen and J. Dean, "Efficient estimation of word representations in vector space," Jan. 2013. [Online]. Available: https://arxiv.org/abs/1301.3781

[25] N. ME, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, 2006.

[26] I. Blekanov, S. Bodrunova, and A. Akhmetov, "Detection of hidden communities in twitter discussions of varying volumes," *Future Internet*, vol. 13, no. 11, 2021. [Online]. Available: https://www.mdpi.com/1999-5903/13/11/295

[27] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, Feb. 2004. [Online]. Available: https://doi.org/10.1103/physreve.69.026113

[28] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, Oct. 2008. [Online]. Available: https://doi.org/10.1088_F1742-54682F10.10008

[29] N. Dugué and A. Perez, "Directed Louvain : maximizing modularity in directed networks," Université d'Orléans, Research Report, Nov. 2015, please cite the following published version: https://doi.org/10.1016/j.physa.2022.127798 rather than this one. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01231784

[30] E. A. Leicht and M. E. J. Newman, "Community structure in directed networks," *Physical Review Letters*, vol. 100, no. 11, Mar. 2008. [Online]. Available: https://doi.org/10.1103/physrevlett.100.118703

[31] V. A. Traag, L. Waltman, and N. J. van Eck, "From Louvain to Leiden: guaranteeing well-connected communities," *CoRR*, vol. abs/1810.08473, 2018. [Online]. Available: http://arxiv.org/abs/1810.08473

[32] S. Fortunato and M. Barthélemy, "Resolution limit in community detection," *Proceedings of the National Academy of Sciences*, vol. 104, no. 1, pp. 36–41, 2007. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.0605965104

[33] V. A. Traag, P. Van Dooren, and Y. Nesterov, "Narrow scope for resolution-limit-free community detection," *Phys. Rev. E*, vol. 84, p. 016114, July 2011. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.84.016114

[34] M. Rosvall, D. Axelsson, and C. T. Bergstrom, "The map equation," *The European Physical Journal Special Topics*, vol. 178, no. 1, pp. 13–23, Nov. 2009.

[35] D. A. Huffman, "A method for the construction of minimum-redundancy codes," *Proceedings of the IRE*, vol. 40, no. 9, pp. 1098–1101, 1952.

[36] J. Benton, "Most people on twitter don't live in political echo chambers - but mostly because they don't care enough to bother building one," Oct 2022. [Online]. Available: https://www.niemanlab.org/2022/10/most-people-on-twitter-dont-live-in-political-echo-chambers-but-mostly-because-they-dont-care-enough-to-bother-building-one/

[37] R. F. Betzel, "Community detection in network neuroscience," 2020. [Online]. Available: https://arxiv.org/abs/2011.06723

[38] A. Lancichinetti and S. Fortunato, "Consensus clustering in complex networks," *Scientific Reports*, vol. 2, 2012.

[39] A. Tandon, A. Albeshri, V. Thayananthan, W. Alhalabi, and S. Fortunato, "Fast consensus clustering in complex networks," *Physical Review E*, vol. 99, no. 4, Apr. 2019. [Online]. Available: https://doi.org/10.1103_physreve.99.042301

[40] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[41] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python natural language processing toolkit for many human languages," 2020.

[42] M. F. Lars Borin and J. Roxendal, "Korp – the corpus infrastructure of Språkbanken," ser. Proceedings of LREC, 2012, pp. 474–478. [Online]. Available: https://spraakbanken.gu.se/verktyg/korp

[43] Y. Tsvetkov, M. Faruqui, W. Ling, G. Lample, and C. Dyer, "Evaluation of word vector representations by subspace alignment," in *Proc. of EMNLP*, 2015.

[44] L. Borin, M. Forsberg, and L. Lönngren, "Saldo: a touch of yin to wordnet's yang," *Language Resources and Evaluation*, vol. 47, no. 4, 2013.

[45] S. Hengchen and N. Tahmasebi, "Supersim: a test set for word similarity and relatedness in Swedish," in *In Proceedings of the 23rd Nordic Conference on Computational Linguistics*, 2021.

[46] T. P. Adewumi, F. Liwicki, and M. Liwicki, "Corpora compared: The case of the Swedish Gigaword & Wikipedia corpora," *CoRR*, vol. abs/2011.03281, 2020. [Online]. Available: https://arxiv.org/abs/2011.03281

[47] R. Rehurek and P. Sojka, "Gensim–python framework for vector space modelling," *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, vol. 3, no. 2, 2011.

[48] J. A. Hartigan and P. M. Hartigan, "The Dip Test of Unimodality," *The Annals of Statistics*, vol. 13, no. 1, pp. 70 – 84, 1985. [Online]. Available: https://doi.org/10.1214/aos/1176346577

[49] F. Huszár, S. I. Ktena, C. O'Brien, L. Belli, A. Schlaikjer, and M. Hardt, "Algorithmic amplification of politics on Twitter," *Proceedings of the National Academy of Sciences*, vol.

119, no. 1, p. e2025334119, 2022. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2025334119

[50] W. Cota, S. C. Ferreira, R. Pastor-Satorras, and M. Starnini, "Quantifying echo chamber effects in information spreading over political communication networks," *EPJ Data Science*, vol. 8, 2019.

[51] J. Shore, J. Baek, and C. Dellarocas, "Network structure and patterns of information diversity on Twitter," *MIS Quarterly*, no. 3, pp. 849–927, 2018.

[52] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *The Journal of Machine Learning Research*, vol. 3, 2003.

[53] D. Colla, M. Delsanto, M. Agosto, B. Vitiello, and D. P. Radicioni, "Semantic coherence markers: The contribution of perplexity metrics," *Artificial Intelligence in Medicine*, vol. 134, p. 102393, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365722001440

[54] F. Morstatter and H. Liu, "In search of coherence and consensus: Measuring the interpretability of statistical topics," *J. Mach. Learn. Res.*, vol. 18, no. 1, p. 6177–6208, Jan. 2017.

[55] M. E. Roberts, B. M. Stewart, and D. Tingley, "STM: An R package for structural topic models," *Journal of Statistical Software*, vol. 91, no. 2, p. 1–40, 2019. [Online]. Available: https://www.jstatsoft.org/index.php/jss/article/view/v091i02

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERNG
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY