Nuclear (e)motives: The KCNA's English language output and sentiment patterns regarding the topic of nuclear weapons

published by Clayton Besaw on Aug 7, 2020

In an age of ever proliferating data and diverging geo-political interests, government statements have become increasingly important source material for understanding the potential communicative logic of political leaders.

Breaking down North Korea's official Korean Central News Agency (KCNA) output in English, we find a clear bifurcation in the pattern of language used when it comes to the discussion of texts related to the topic of nuclear weapons.

In a collection of news articles with a positive sentiment, words suggest an association with positive domestic/political benefits or weapon disarmament. In contrast, the collection of news articles with a negative sentiment more distinctly highlight and focus on hostile language and foreign threat perception.

This analysis outlines the methodologies we used to come to these conclusions and provide the reader with visual evidence regarding the affective patterns that manifest within the KCNA's English language output.

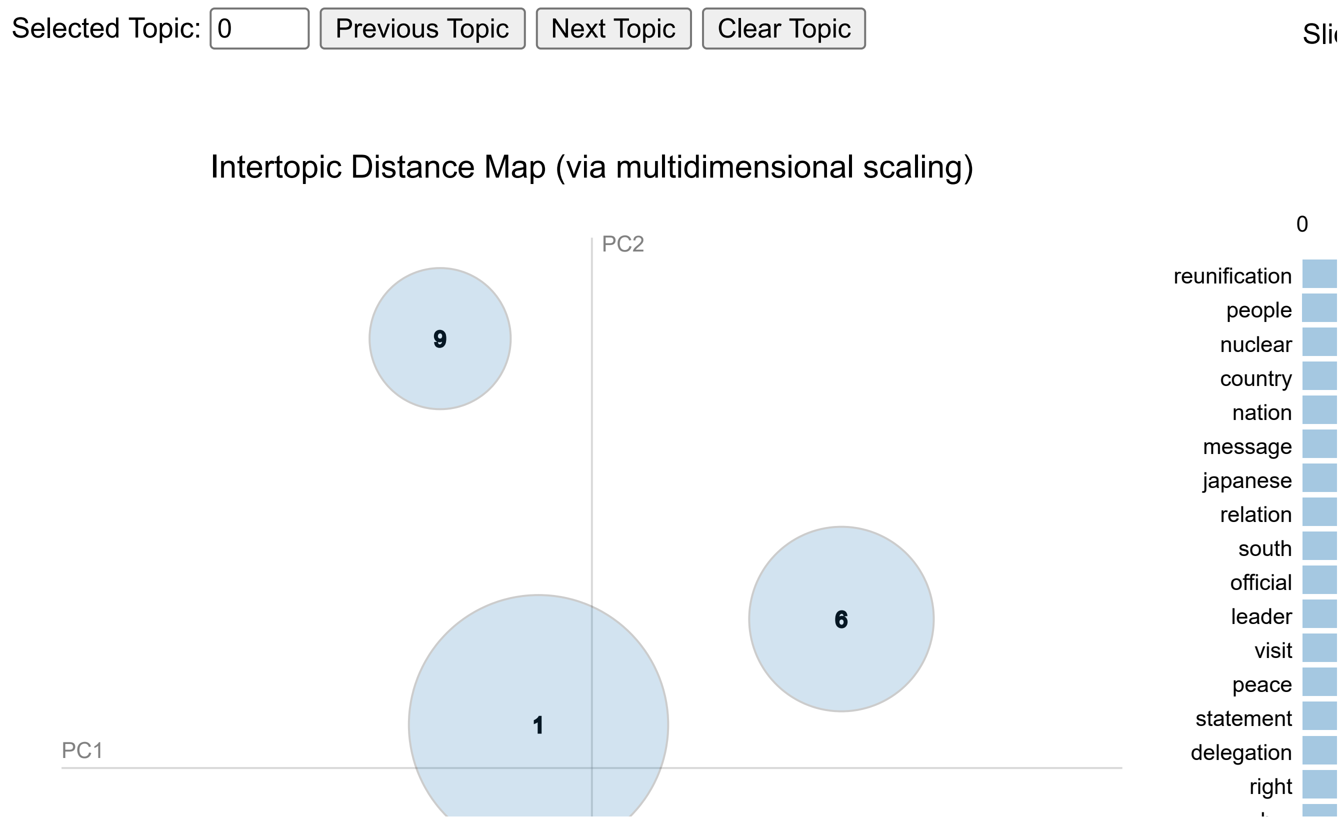
Analytical Introduction

To get to our goal of exploring sentiment patterns as related to the KCNA's discussion of nuclear topics, we first turn to topic modeling.

Topic modeling, as a form of computational OSINT, is useful for us because it allows for a process of quantifying the interrelationships between corpus topics and how their presence changes over time. [1]

However, this analysis will attempt to dig a bit deeper into our KCNA corpus going back to 1996 by utilising the insights from our topic modeling tool and well-known sentiment libraries.

While topic modeling can provide us with a powerful tool for exploring broad topics within a large amount of text data, it does not necessarily tell us much about sentiment-patterns within topics.



KCNA English Langage Corpus Topic Modeling Dashboard

When we trained our topic modeling algorithm on the KCNA text corpus, we found ten distinct topics that manifest across the texts, based on the word clusters associated them, we named them: (1) Juche/Ideology, (2) Nuclear, (3) Science/Tech, (4) Negative ROK, (5) Historical Grievance, (6) Positive ROK, (7) Military/Patriotism, (8) Art/Sport, (9) Diplomacy and (10) Foreign

Negative ROK, (5) Historical Grievance, (6) Positive ROK, (7) Military/Patriotism, (8) Art/Sport, (9) Diplomacy and (10) Foreign Relations.

The topics are also ordered in their respective volumes from largest (Juche) to smallest (Foreign Relations). Volume here meaning the percentage of the overall KCNA English corpus that consists of words associated with each topic.

The right hand side of the topic modeling dashboard above provides a list of the 30 most salient words associated with each topic, with these words helping us to designate subjective labels to each topic. [2]

While label names such as "Juche/Ideology" or "Nuclear" may have their own implicit subjective connotations when it comes to interpretation, there are likely to be meaningful affective patterns within each topic that would be of interest to an analyst of North Korean state media.

Given that Datayo and Open Nuclear Network's primary focus is on nuclear risk reduction, we thus sought to uncover any potential meaningful patterns within sentiment across KCNA documents that focus on our previously uncovered "Nuclear" topic.

We will first examine sentiment across the overall KCNA corpus from 1996 to 2020. Second, we will look at inter-topic variation across sentiment. Finally, we will explore intra-topic patterns in sentiment within a set of nuclear-related KCNA articles.

Sentiment in the KCNA's English Language Output (1996 - 2020)

To get a better sense of how sentiment manifests within nuclear-related KCNA documents, it is prudent to first examine how sentiment manifests across the entire KCNA corpus. Each article published during this time is treated as a unique document.

In calculating sentiment, we utilised our corpus of KCNA English language documents alongside the TextBlob library [3] in Python.

TextBlob utilises a sentiment library known as the The Pattern Library for calculating sentiment (also called polarity). Sentiment ranges from -1 (most negative) to +1 (most positive) and individual word scores are based on the Pattern Library's expert coding of word sentiment. TextBlob will find words in our text documents and calculate sentiment for each before averaging them for the entire document of interest. [4]

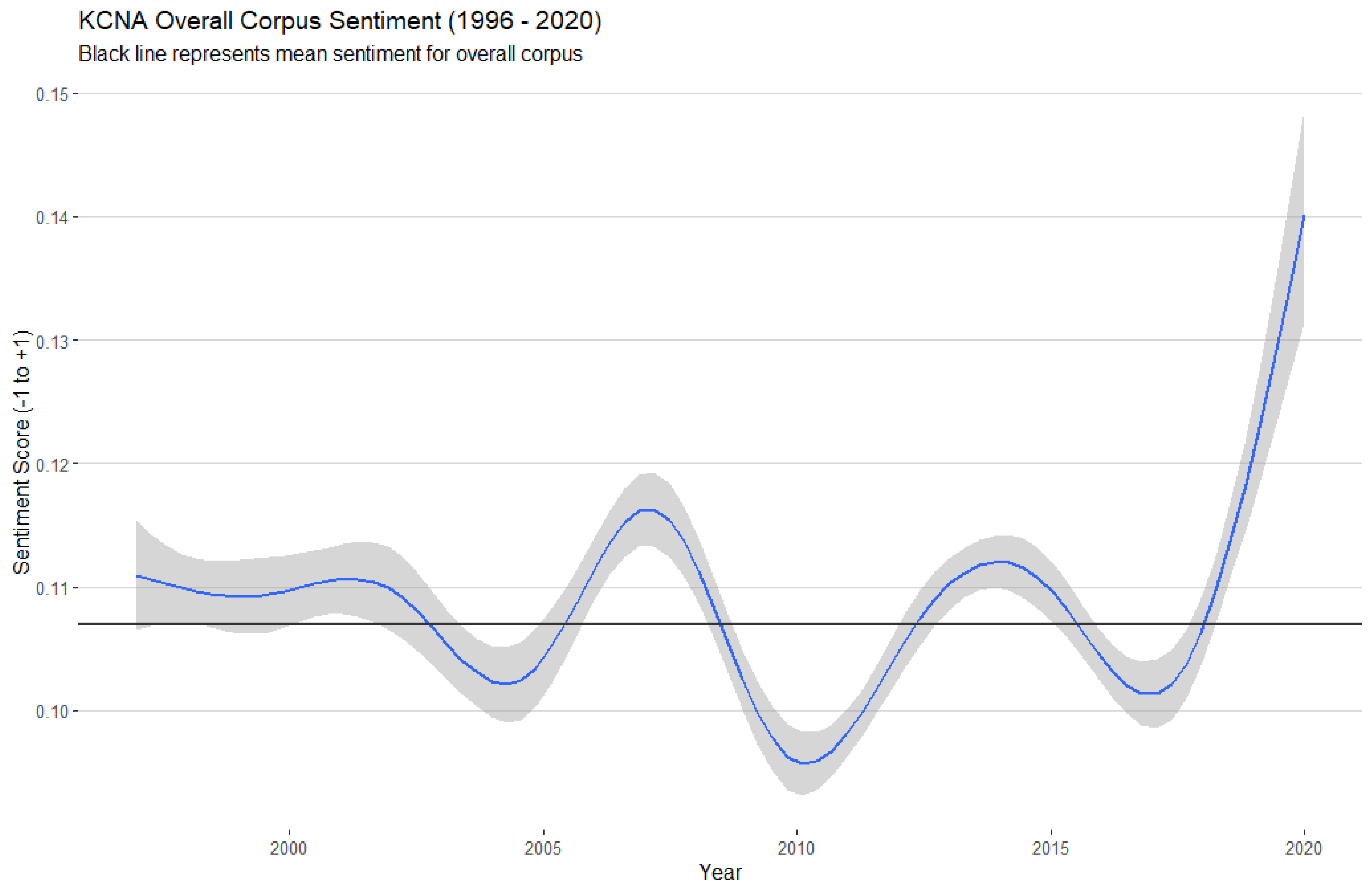


Figure 1

So how does our KCNA corpus look through the lens of sentiment?

Overall, the corpus trends slightly positive with a mean sentiment of 0.107 (Black horizontal line in figure 1).

However, this is also quite close to 0, which suggests that the central tendency pushes towards neutral/low positivity across the KCNA's English language output.

Figure 1 visualises smoothed time-series trends for sentiment by year. The black line represents our overall corpus mean of 0.107 and the smoothed yearly sentiment scores give us an indicator for how the corpus has deviated from the central tendency across the last 25 years or so.

In examining these trends, we see two key patterns emerge.

First, the average yearly sentiment ranges from a low of 0.083 in 2011 to a high of 0.151 in 2020 (so far!). With an estimated increase in positive sentiment starting in 2018.

Second, the yearly average seems to be modestly volatile across our collection period with numerous dips and gains relative to the overall central tendency. Relatively negative periods occur between 2002 to 2005 (collapse of Agreed Framework in 2002 until Six-Party Talks breakthrough in September 2005 Joint Statement), a year from 2010 to 2011 (Cheonan and Yeonpyeong Island incident in 2010 until death of Kim Jong Il and succession by Kim Jong Un in 2011), and also from 2016 to 2017 (high levels in US-DPRK hostilities amid three nuclear tests, ICBM tests, additional UNSC sanctions and highly bellicose language exchange).

The takeaway is that the average yearly sentiment of the KCNA's English language output trends slightly positive with regular shifts above and below the central tendency of 0.107.

So what happens when we break this down by topic?

Inter-Topic Sentiment

For this analysis we will focus on the top-4 topics in terms of representation across the corpus that contain significant sentiment variation.

These topics are Juche/Ideology, Nuclear, Negative ROK and Historical Grievance. While Science/Tech is the third biggest topic in terms of volume, it has primarily a neutral sentiment with no significant subset of documents that have a strong positive or negative sentiment.

For the visual and statistical comparisons, we make a subset of documents that are most closely related to these topics.

Thankfully, our topic modeling algorithm provided an effective way to achieve this. When our topic modeling algorithm was trained on the KCNA corpus, it provided us with an estimated proportion for which that document reflects our latent topic clusterings. For each topic, we subset the documents that have at least a 51% clustering of words associated with each topic.

Examining central tendency first, we find some interesting patterns in the manifestation of sentiment across these four topics.

For this analysis we will focus on the top-4 topics in terms of representation across the corpus that contain significant sentiment variation.

These topics are Juche/Ideology, Nuclear, Negative ROK and Historical Grievance. While Science/Tech is the third biggest topic in terms of volume, it has primarily a neutral sentiment with no significant subset of documents that have a strong positive or negative sentiment.

To further compare sentiment patterns across these topics, we first need to make a subset of documents that are most closely related to these topics.

Thankfully our topic modeling algorithm provided an effective way to achieve this.

When our topic modeling algorithm was trained on the KCNA corpus, it provided us with an estimated proportion for which that document reflects our latent topic clusterings. For each topic, we subset the documents that have at least a 51% clustering of words associated with each topic.

Examining central tendency first, we find some interesting patterns in the manifestation of sentiment across these four topics.

Juche/Ideology has an average positive score with a mean sentiment of 0.24. In contrast, Nuclear, Negative ROK and Historical Grievance are on average slightly negative with a mean sentiment of -0.015, -0.0004 and -0.013 respectively.

Substantively, this is not all that surprising. Documents associated with the regime's ideology are likely to be positive by virtue of their role in cultivating support for the government. In contrast, the other topics seem to more primarily couched in negative affective intent.

The above observation also seems to manifest when more rigorously examining each topic's statistical qualities.

Figure 2 below shows a box plot of for each topic. A box plot is a method for graphically depicting group differences across a quantitative indicator based on median (black line within box) and quartile differences (top and bottom edges of box).

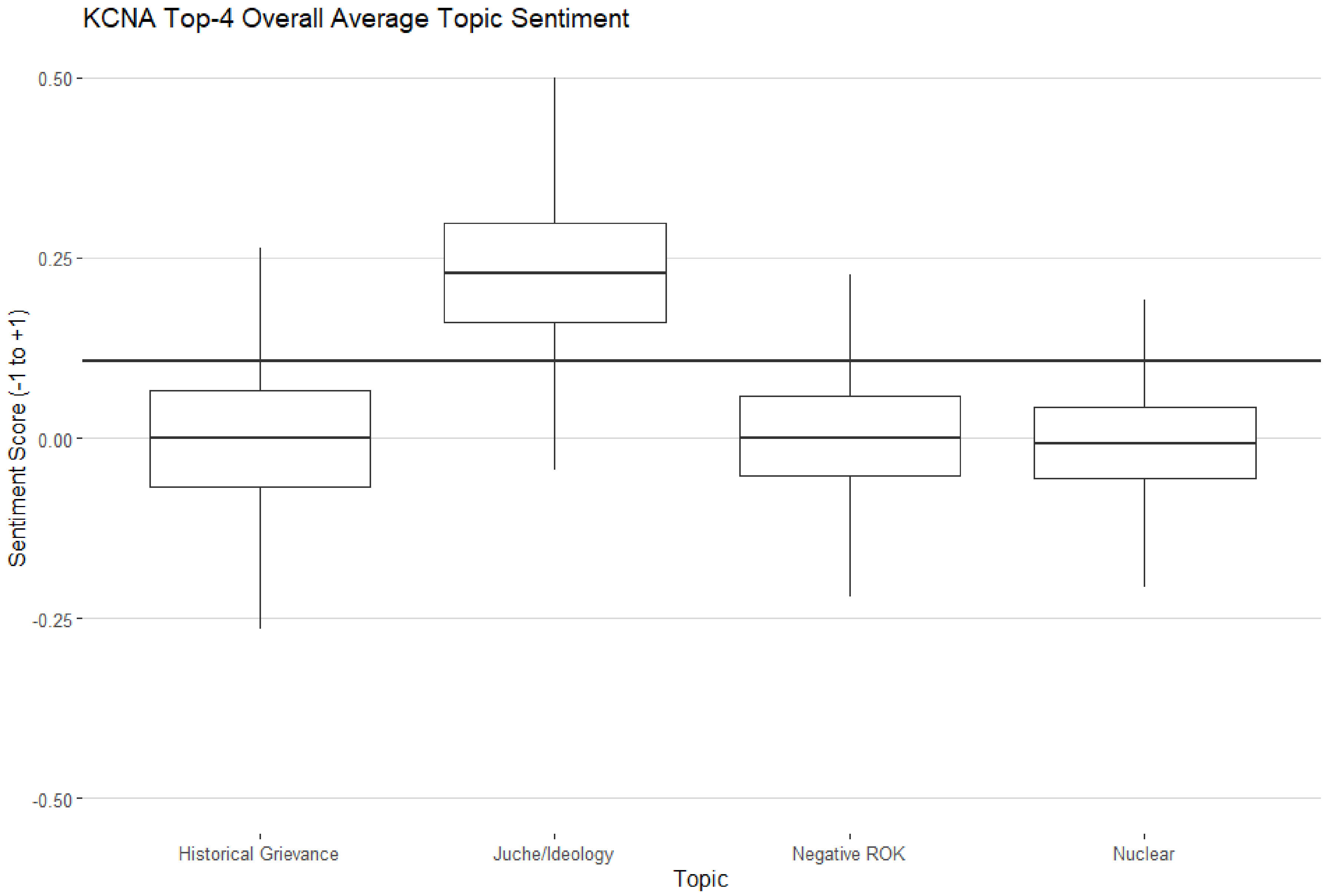


Figure 2

Looking at median sentiment by topic, Juche/Ideology is clearly above the overall corpus mean of .107. As expected, Nuclear, Negative ROK and Historical Grievance have mean topic values that fall well below the overall corpus mean.

All four topics have considerable variation, however, meaning that positive and negative sentiments exist within documents found in each topic clustering.

Looking at the box-plot in figure 2 shows that each topic's minimum and maximum (vertical lines above and below the quartile range) contains both negative and positive sentiment values respectively.

To get a sense of how these quantities might have changed overtime, figure 3 below shows a smoothed time-series comparison between all four topics.

Juche/Ideology remains mostly consistent over time and is well above the average corpus sentiment.

Historical Grievance, Negative ROK and Nuclear all remain consistently below the average corpus sentiment and often straddle the line between neutral (0) and negative (< 0) sentiment across the corpus time period.

Interestingly, the Historical Grievance and Nuclear topics track in an almost collinear fashion across the corpus time period.

This is not substantively surprising given that we will soon learn that the regime's nuclear ambitions are often couched in language tied to threat perception and military deterrence vis-a-vis historically perceived hostile actors.

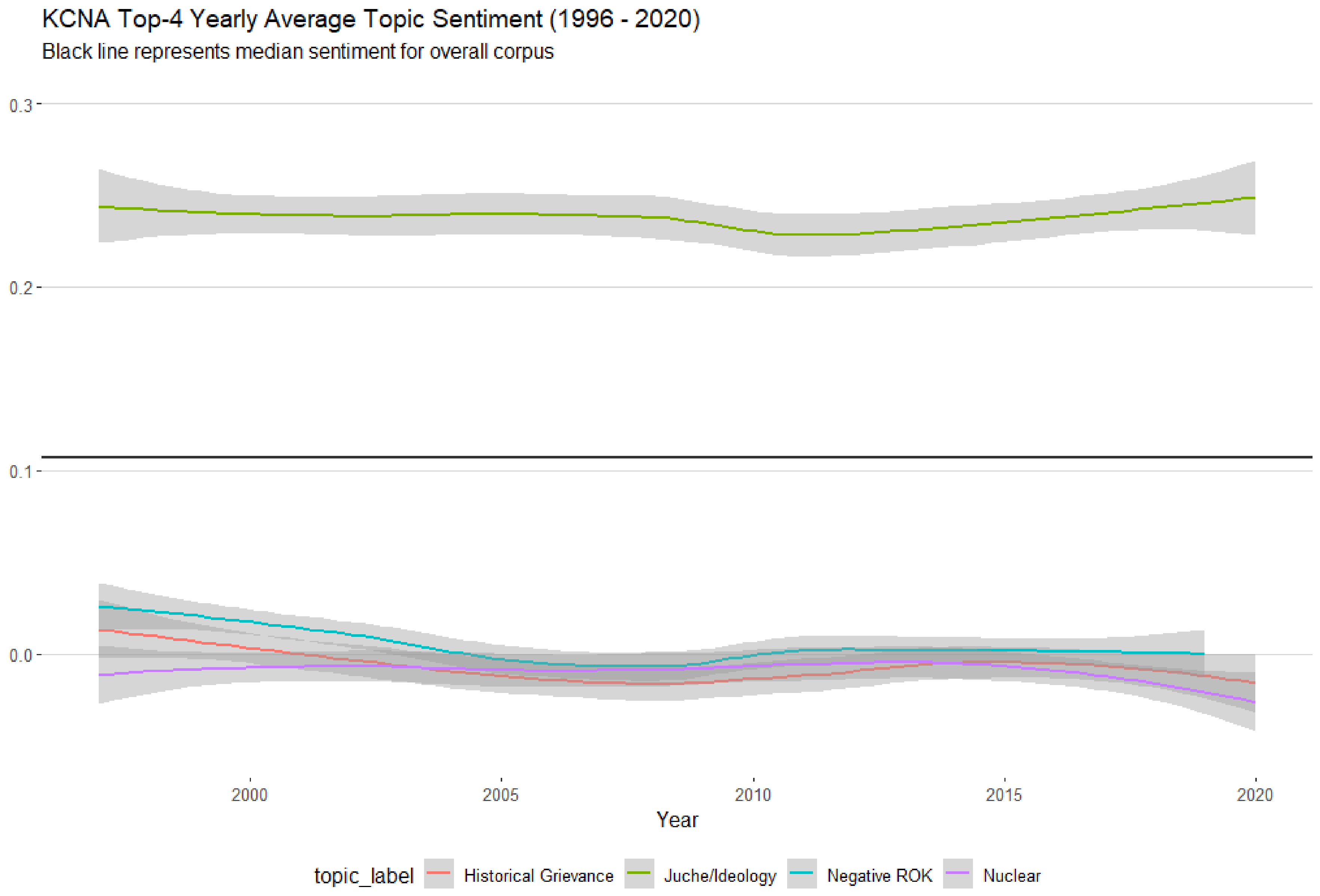


Figure 3

So what have we learned here?

First, there appears to be considerable variation across topics in how sentiment manifests both on average and across time.

Second, our topic of interest (nuclear) appears to be consistently negative on average. In addition, the nuclear topic also tracks very linearly with Historical Grievance in terms of changes in sentiment.

Finally, our boxplot shows clear evidence for the fact that there might be some interesting variance within topics themselves. Even if Nuclear is on average a negative topic in terms of sentiment, it is possible that there exists some meaningful variation to unpack within the topic itself.

Positive and Negative Sentiment within the Nuclear Topic

A first step to unpacking sentiment variation within our Nuclear topic sub-corpus is to categorise each document as positive, negative and neutral respectively.

As stated in the beginning of this analysis, we can do this by categorising a document with a sentiment above zero as positive, below zero as negative and at zero as neutral. Figure 4 shows a bar chart of the sentiment categories and their relative frequencies within the subset of documents related to the Nuclear topic.

Overall, there are 8287 documents that can be classified as being associated with our Nuclear topic clustering.

Roughly 57% of these Nuclear documents can be categorised as having a negative sentiment. 42% can be categorised as having a positive sentiment and roughly 1% can be categorised as having a neutral sentiment.

This suggests that the Nuclear topic has a clear skews towards negative sentiment manifestation but that it roughly stays between the negative-positive dichotomy with very few “neutral” documents.

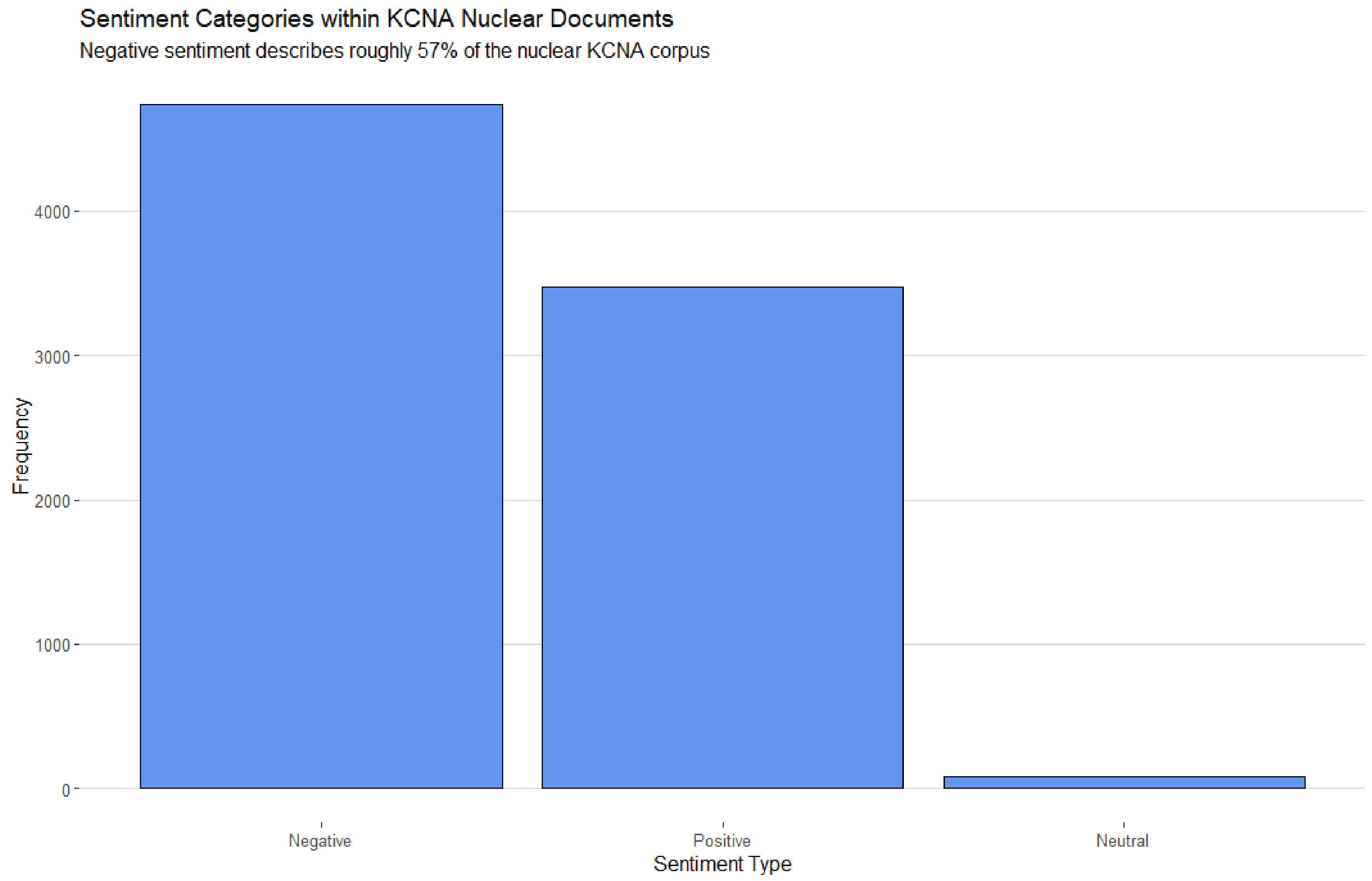


Figure 4

So what makes a negative or positive Nuclear topic document in this case?

To dig a little deeper, we can plot the most common words found within our negative and positive nuclear documents respectively. We can also see which words are most statistically associated with the specific term "nuclear" in both the positive and negative sub-corpora as well.

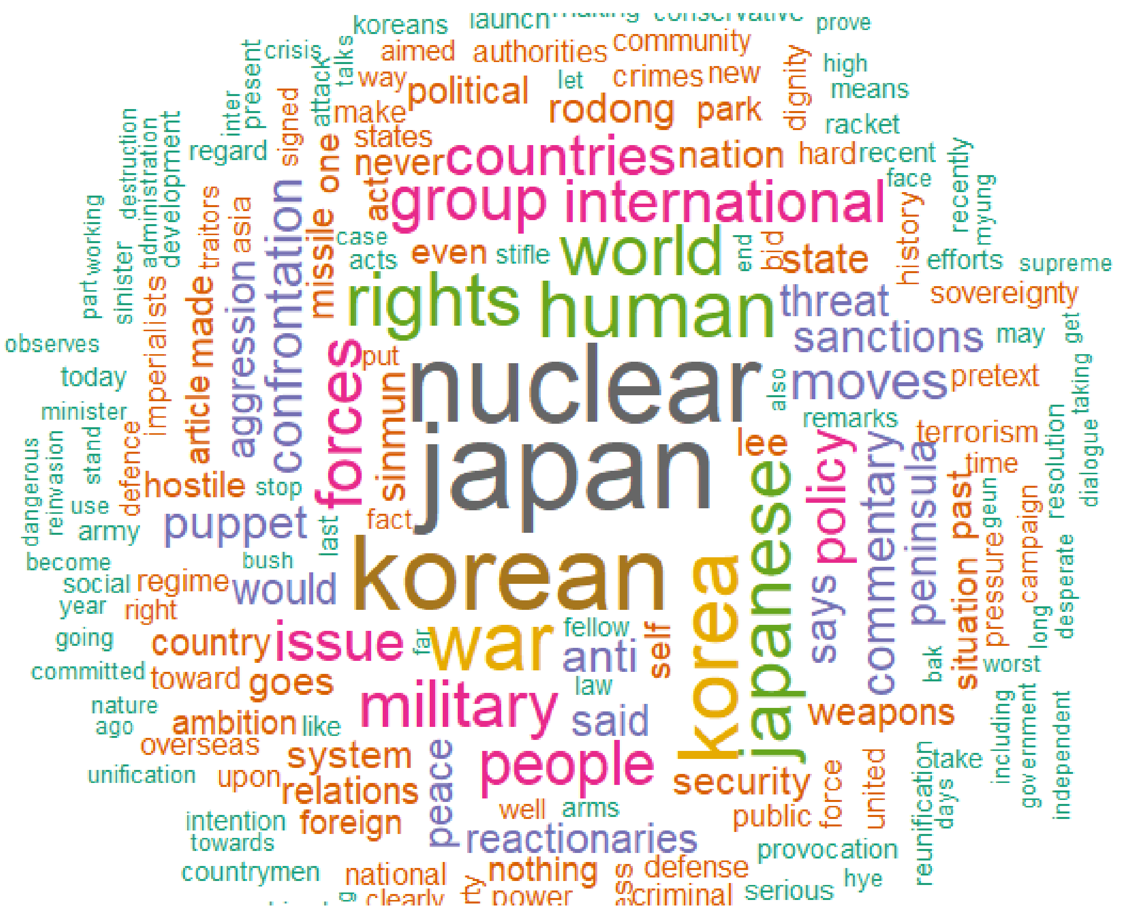


Figure 5: Negative Sentiment Nuclear Documents Top 200 Word cloud

Figure 5: Negative Sentiment Nuclear Documents Top-200 Wordcloud

First examining the negative Nuclear corpus, we find a few unsurprising patterns. Figure 5 shows a word cloud for the top-200 words within the negative corpus and prominently displays words such as “japan”, “war”, “military”, “threat”, “sanctions”, “confrontation”, “puppet”, “terrorism” and “aggression”.

This suggests that the negative Nuclear corpus is centered around words that represent foreign actors and conspicuously hostile language.

When examining the positive Nuclear corpus, we find some surprising overlap and some unsurprising differences. Figure 6 displays the respective word cloud for the top-200 words within the positive corpus.

Terms such as “japan” and “nuclear” are still the most prominent words even within the positive corpus. Even words such as “war”, “sanctions” and “military” make a prominent appearance. However, there are some interesting differences when you look at how the rest of the frequent terms manifest.

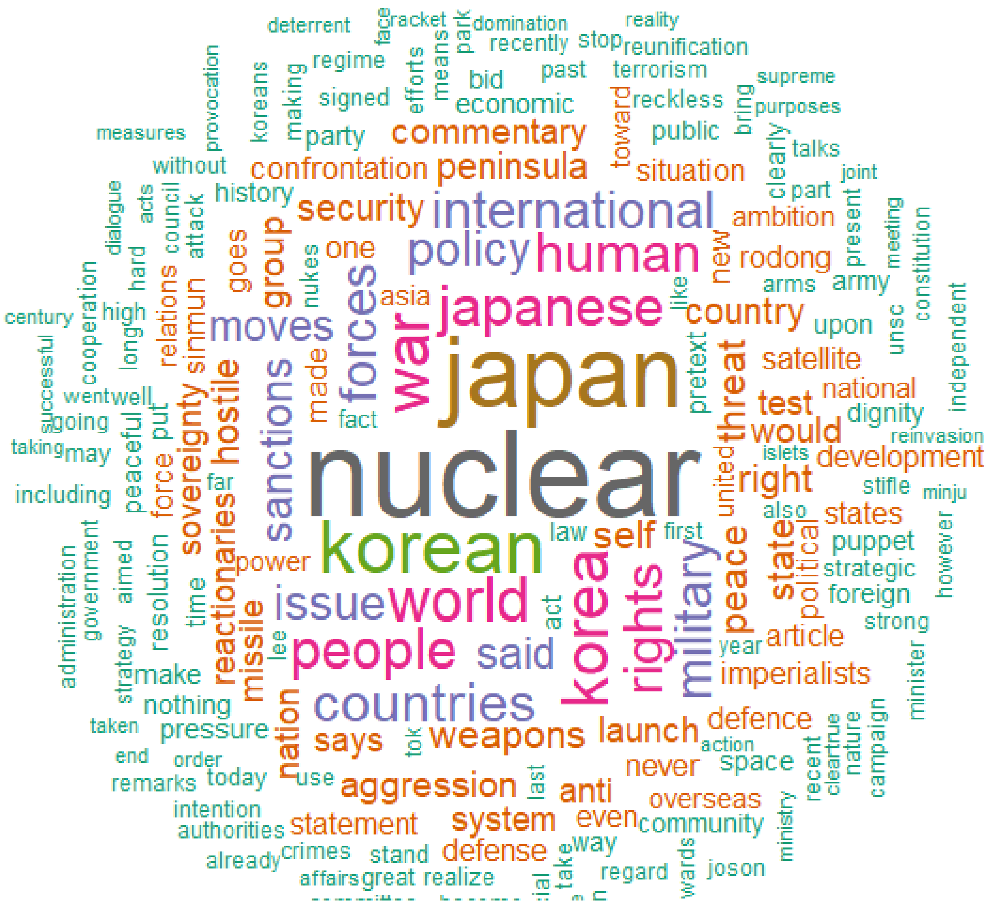


Figure 6: Positive Sentiment Nuclear Documents Top-200 Wordcloud

For example, there are more words related to domestic political actors and security issues. “people,” “army,” “security,” “launch,” “test,” “development,” “sovereignty,” “self” and even “peace” show up as some of the most frequent terms used within the positive sentiment corpus.

Overall, this suggests that while the most frequent terms used overlap quite heavily between positive and negative Nuclear documents, the bulk of the most frequent terms used seems to have some divergence.

Namely that the negative Nuclear corpus is more centered on foreign actors and their perceived threat. Associating hostile motives to others and implicitly painting the country as a victim of these hostile foreign threats.

In contrast, the positive corpus brings in words that focus more on North Korea itself. Framing language around ideas such as development, security, state sovereignty and peace. The implication here is that the more positive Nuclear documents are in sentiment, the more likely they will touch on domestic justification and deterrence.

Table 1: Statistically associated words with the term "nuclear" in the negative and positive Nuclear corpora

Sentiment Type	Words (Correlation Coefficient)
Negative	weapons (0.63), nukes (0.42), peninsula (0.37), access (0.35), threat (0.33), deterrent (0.32), blackmail (0.32), proliferation (0.32)
Positive	weapons (0.70), nukes (0.49), access (0.43), disarmament (0.39), arsenal (0.37), deterrent (0.36), proliferation (0.35), threat (0.33), modernization (0.33), blackmail (0.31), tests

modernization (0.33), blackman (0.31), tests (0.30), warheads (0.30)

We can find further information about the differences between the positive and negative Nuclear corpora by examining the statistical associations of other words with the specific term "nuclear." Table 1 displays the most correlated words with "nuclear" in both sets of documents by using a Pearson correlation coefficient that ranges from -1 (negative correlation) to +1 (positive correlation).

As before with our word clouds there is considerable overlap between both corpora, but also clear differences.

In the negative corpus, nuclear is most commonly correlated with words associated with threat perception. In contrast, the positive corpus also includes less "threat-laden" words such as disarmament and modernisation. The positive corpus also speaks more to domestic development ideas such as testing and warhead development.

Taken in total, it is clear that the main difference between the negative and positive Nuclear corpora is that the positive corpus highlights words that speak more to the positive domestic benefits of nuclear weapons possession ("deterrent", "modernization", "access") while also including more constructive terms such as "disarmament". In contrast, the negative corpus is almost entirely focused on foreign threat perception.

Given this finding, we can close this analysis by plotting the average yearly sentiment for both our positive and negative corpora against each other. We know now that there are clear word choice differences between the two and it would be of interest to us to see how these sub-corpora change in relation to each other over time.

Figure 7 displays the time-series comparison in terms of yearly average sentiment for both our negative and positive corpora.

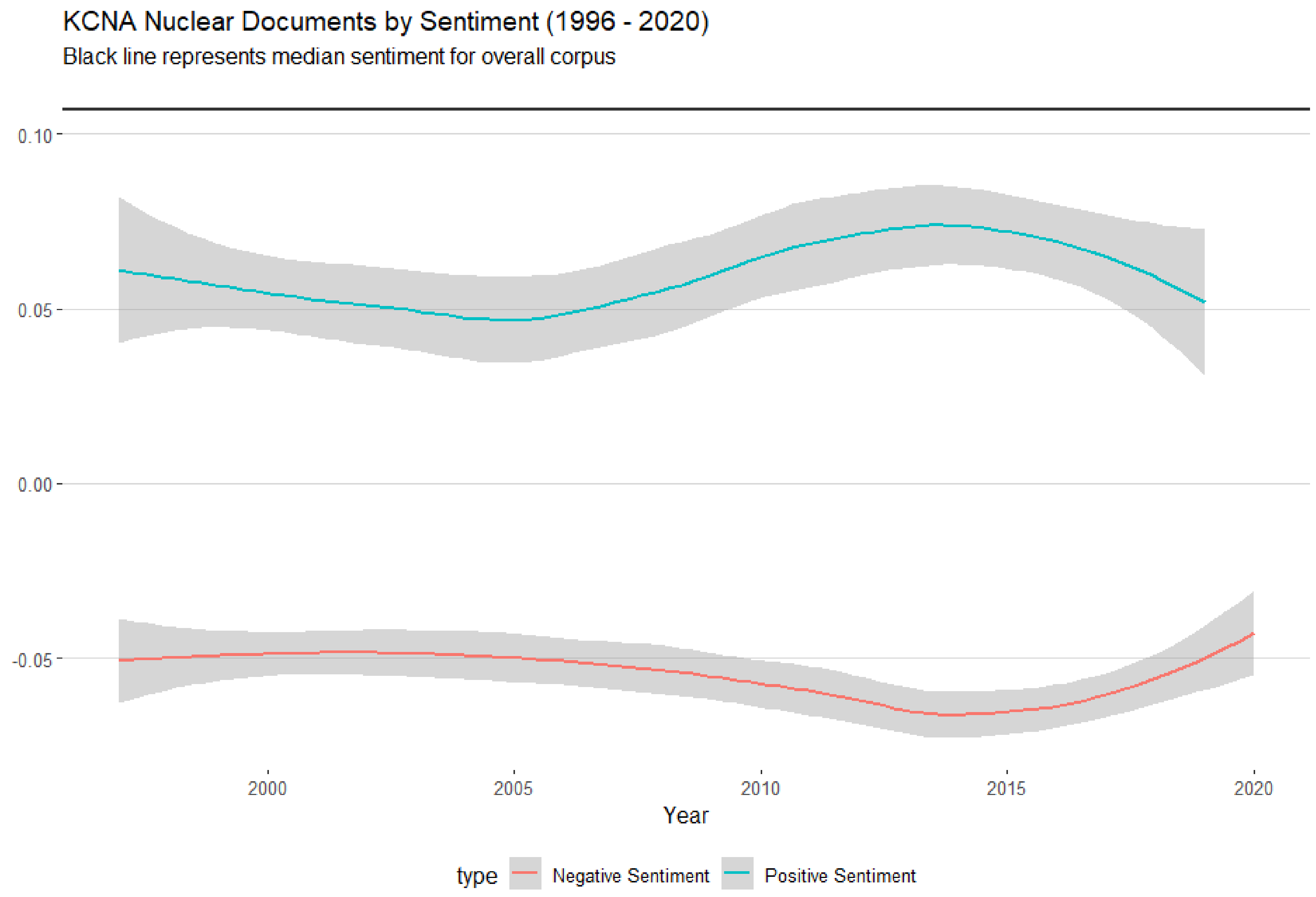


Figure 7

First, both corpora are well below the average overall corpus sentiment of 0.107, which means that even the positive sentiment sub-corpus is relatively more negative than the average KCNA news article.

Second, there appears an almost inverse correlation between the two in the post 2005 period.

As the average positive sentiment in the positive corpus began to increase, the average negative sentiment in the negative corpus saw a modest decrease in tandem. Increased positive sentiment also came with increased negative sentiment. In 2015, both the positive corpus became less positive while the negative corpus became less negative.

It is noteworthy that these developments followed North Korea's first nuclear test in 2006 which appears to have elicited

It is noteworthy that these developments followed North Korea's first nuclear test in 2006 which appears to have elicited more intensified language in its nuclear-related media output. After a peak that was more pronounced in the positive corpus, state media shifted towards a more moderate tone, reaching pre-first nuclear test levels in 2019. This moderation in the nuclear corpus occurred despite the highly fluctuating years of 2016-2019 in terms of foreign relations with its neighbours and the United States.

While it is incredibly difficult to unpack why this is happening from such a superficial visualisation, it does suggest that increases in the positive sentiment related to KCNA Nuclear documents tend to go hand in hand with increases in negative sentiment. Essentially highlighting both aspects of this negative (foreign threats) and positive (self development and prestige) simultaneously.

[1] In a previous article concerning Russian presidential speeches here on Datayo, we explored the utility of topic modeling for the extraction of meaningful concept clusters in large sets of text documents. [link] Since that analysis, we have released a new text exploration tool on Datayo. So far we have built a dashboard for exploring topics, and their associated words, within the entire corpus and for visualizing each topic across time. This tool is currently available for our regularly updated KCNA English corpus, but will soon cover official Russian presidential speeches and eventually official political documents from all six parties to the Six-Party Talks negotiations concerning the denuclearisation of the Korean Peninsula.

[2] We suggest taking some time to explore the individual words that make up each topic within the corpus. A conservative lambda/relevance cut-off of .4 (upper right-hand slider) is a good metric for examining the kinds of words that are most relevant to each topic as determined by our topic modeling algorithm.

[3] https://textblob.readthedocs.io/en/dev/

[4] For a more rigorous detailing of how specific sentiment is calculated with the inclusion of word modifiers, see this excellent breakdown by Aaron Schumacher of Planspace. https://planspace.org/20150607-textblob_sentiment/

Machine Learning Natural Language Processing Political Speech DPRK KCNA Sentiment Nuclear Language Threat Perception

Share this on :