Patterns and Biases in International News Flow: A Machine Learning Approach

Mickey Fraanje
STUDENT NUMBER: 2006951

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
BACHELOR OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:

dr. H. J. Brighton
dr. E. Keuleers

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
June 2020

Preface

This thesis was written to fulfill the graduation requirements for the bachelor Cognitive Science and Artificial Intelligence at Tilburg University. Working on this project has taught me more than I could imagine. It was a valuable and interdisciplinary challenge that led to significant improvements in my coding skills, data science knowledge and research capabilities and I feel it has granted me a deeper understanding of the field and the techniques that I have been studying for the past few years. Of course, it would not have been possible to complete this project without the support of the following people. Firstly, I would like to thank my supervisor dr. Henry Brighton for his excellent guidance and advice. Each weekly meeting gave new, valuable perspectives to help with the progression of this project. I also want to thank Sander and Ronja. Being able to share thoughts, ideas and laughs proved to be very motivating while working from home during these unconventional times. And finally, my sincere thanks to Lies, my friends and my family for all of their invaluable support and encouragement.

Mickey Fraanje
Breda, June 2020

Patterns and Biases in International News Flow: A Machine Learning Approach

Mickey Fraanje

There has been an interest in bringing the hidden structures of international news flow to light for some time. However, the main focus of research has thus far been on finding the determinants and statistical relationships that underlie news coverage patterns. The current study proposes a modern, data-driven approach to exploring this subject which may allow these patterns and biases to be found automatically using both supervised and unsupervised machine learning techniques. It was found that countries displays distinct characteristics in their style of news coverage by which they can be identified using machine learning. Not only by observing which countries they cover but also by how they cover them. Using sentiment data to classify countries appears to be less accurate than using their references to other countries but the fact that it is possible shows interesting promise for future research. In the second experiment, which made use of clustering techniques, results show that geographical proximity is still the primary factor behind news flow. In addition, the visualisation of this process gives further insight into the specific relationships between regions. Finally, not only do specific countries display unique patterns, national characteristics such as human development index, GDP growth and population growth also seem to have distinct patterns associated with them to some degree. This could provide potential new methods of predicting these variables in future research.

1. Introduction

The highly interconnected nature of modern global society means that people's perceptions of the world around them is more important than ever before. However, since not everyone has the means to travel the world or the motivation to extensively research foreign lands, the average person's worldview is still largely dependent on the way their home country covers international news. For instance, in their study about the agenda setting effects of news coverage, Wanta, Golan, and Lee (2004) state that the amount to which another country is covered in the news has a significant impact on the public's perceived importance of this country in regards to their own national interests. This also includes the possibility of creating a negative image of the nation in question if the news uses a more negative sentiment in its coverage. In a democratic civilisation especially, where the public shapes the way the country is governed, the underlying biases of a system that causes potentially distorted images of foreign nations is a very relevant topic indeed.

The hidden structures that form the way international news is covered and the determinants behind this system have captured the interest of researchers for quite some time (Kariel and Rosenvall 1984; Chang, Shoemaker, and Brendlinger 1987). The previously mentioned studies have mainly focused on identifying variables, determinants and examining their effects on global news coverage with traditional statistical methods. However, the utilisation of machine learning models on a globally compre-

hensive data set such as GDELT (Global Database of Events, Language and Tone) allows for the use of a novel bottom-up approach to the subject instead of the top-down perspective that has been prevalent so far. In other words, letting the data speak for itself. This could potentially lead to new and interesting insights that ultimately give a deeper understanding of country's biases towards each other on a global scale, solely using news data.

The automated, big data oriented nature of this method could allow for a more efficient method of researching the topic, test hypotheses and provide alternative perspectives on previous findings. For example, regionalism has been found to be one of the most significant factors in regards to news flow (Kwak and An 2014). What if this factor is eliminated in its entirety? Will countries still display distinctive patterns and biases? If using GDELT and machine learning proves to be an effective method, questions like these can be answered with ease and efficiency. Using this methodology, this study will account for regionalism by simply including sentiment data that is readily available from GDELT during the experiments. This could potentially provide evidence that there are not only distinctive patterns in who certain countries include in their coverage, but also how that coverage is conveyed. Similarly, this approach could bring new applications to the news data that GDELT provides. Perhaps there are shared patterns that characterise countries that have a similar level of GDP growth, population growth or human development.

1.1 Problem Statement and Research Questions

To summarise, this study will explore a novel, data-driven methodology to find out whether this is an effective approach for studying patterns and biases in international news coverage. Furthermore it will use these machine learning techniques to take a look at some potential practical applications that could emerge from it. To this end, the project will use an overarching research question with 3 sub-questions:

Can supervised and unsupervised machine learning techniques be used to effectively explore patterns and biases in international news flow?

1. To what extend can news coverage data and classification algorithms be used to identify countries?
2. Which similarities between countries will become most apparent when clustering countries based on news coverage behaviour using unsupervised learning?
3. Can supervised learning be used to explore news flow patterns that emerge from specific national characteristics instead of specific countries?

1.2 Findings

The results of this study provide some compelling evidence towards the validity of this method. Using the data sets created with GDELT, classification algorithms can be used to identify countries based on the way they cover news with high accuracy. Using data about which other countries they cover gave the best result, as could be expected based on the literature. However, sentiment data could surprisingly be used to this end as well. Though it produced lower results, there is potential to be found in this concept. Especially considering the combination of the two data sets produced the best results. Using clustering techniques with this data indeed reflects geographical similarities. With some interesting exceptions, specific regions of countries would be

clustered together rather accurately. Finally, it proved to be possible to identify national characteristics using this method. Though the models had difficulty with GDP growth, population growth and HDI produced good results. Would this method extend itself to the prediction of national characteristics as well? The results seem to be promising but not conclusive. Further research on this subject is highly encouraged in order to further experiment with the predictive applications of this method.

2. Related Work

The kind of news that is available to a person can have a profound effect on the way they perceive the world around them. As previously mentioned, international news coverage can affect the public's perceived importance of other countries and could potentially paint them in a negative light (Wanta, Golan, and Lee 2004). In addition, Aalberg et al. (2013) found a positive correlation between the amount of international news coverage and how much a person knows about international affairs in general.

Many have researchers have wondered about the underlying structures of international news flow and as a result, much research has been done on the subject. While the idea that countries have different patterns and biases in the way they cover global news would seem like an easy assumption to make, Wilke, Heimprecht, and Cohen (2012) uncovered the concrete differences between nations in their research. For instance, distinctions already arise in the amount of countries featured in a particular country's news coverage. Egypt and Switzerland scored the highest with both of them mentioning a number of 70 countries in the four week period of analysis while at the bottom of the list, the United States featured 36 countries and Japan only 28. However, Egypt and Switzerland differed notably in the ratio of foreign news covered compared to domestic news. Egypt scored 65% and Switzerland 43%. Notably, the United States was found the be the most featured country around the globe followed by the United Kingdom. In regards to continents, Europe is the most discussed. This, together with a collection of other differences found in the study, implies that each country's behaviour in this regard has unique characteristics that could potentially be classified using algorithms.

Kim and Barnett (1996) explored the issue of uncovering the structures of news flow by using a network approach. This included using a clustering analysis to gain insights into the relationships between nations and how they would group together within the network. It was found that the structure of the network could be divided in 8 distinct geographical-linguistic categories: European-North American, Chinese, Portuguese, Greek-Turk, Latin American-Spanish, Middle Asian-Indian, North African and overseas French. This research also suggests the existence of a large imbalance between developing and developed nations in regards to news flow. The age of this research and the use of a newspaper and periodical based data set begs the question, would another analysis in the current digital era, using unsupervised clustering techniques, produce similar results?

2.1 Major Determinants

When it comes to deciding what nations are featured in the news most commonly, several different causes have been suggested. Some of the primary indicators of potential news coverage include, regionalism, trade, whether the country is a dominant superpower and the fact that there appears to be special attention for crisis-ridden regions (Wilke, Heimprecht, and Cohen 2012). The dominant superpower factor especially could explain many of the most covered regions in the world. With both the US

and the UK being two of the world's biggest powers and Europe being the continent with the highest density of superpowers. Alternatively, Segev (2015) states GDP, foreign population and conflict are the most significant variables that determine how often a country will appear in other country's international news. However, not all conflict seems to receive equal amounts of attention since conflict in the Middle-East appears more often in foreign news compared to other crisis-ridden regions. It is hypothesised in the study that this may be the case because of its relative proximity to Europe, which implies that regionalism could potentially be a factor in this case as well.

In regards to regionalism, several studies have found it to be a very, if not the most significant factor when it comes to international news flow (Kwak and An 2014; Wilke, Heimprecht, and Cohen 2012). In essence this means that countries have a bias towards covering news from nearby states and regions. The significant nature of this factor leads to an interesting question. Is it possible to distinguish the news coverage styles of countries once regionalism is eliminated as a variable? The current study will try to answer this by including sentiment data during the experiments. This will hopefully give deeper insights into the patterns of news coverage since it would prove whether there are not only patterns in who certain countries include in their coverage, but also how that coverage is conveyed.

In contrast, Wu (2000) proposes that the two primary determinants are the economic ties between the broadcasting and the featured country as well as the amount of news agencies available in the featured country. Since other research seems to imply that regionalism is one of the the most significant factors, it could be interesting to take a closer look at the relationship between these determinants. Generally speaking, a country's most common trading partner is likely to be their neighbour as well. This could indicate a direct relationship between the variables. Moreover, superpowers are often quite dominant in regards to international trading relationships as well which could prove to be another potential correlation. While the specifics of these questions are outside of the scope of the current study, visually clustering the countries using unsupervised machine learning could give an overview that could potentially provide a deeper insight into this issue.

For the second point, he explains relevance of news agencies by the fact that it is more economical to gather news from a foreign country's own news agency compared to sending a correspondent abroad. While this is a common factor across countries of all development levels, it seems to have a larger influence on the coverage of developed nations (Wu 2003). As previously mentioned, the major reason behind the usage of foreign news agencies is the economical value. Therefore it makes sense that poorer, developing nations make more use of this to gather news on developed nations than vice-versa.

2.2 Developing and Developed Nations

While most determinants are consistent between the news coverage that developed and developing nations receive, there have been findings that indicate even more significant differences between the two types of countries. As has been found in previous research, both trade and the availability of news agencies are significant factors for both groups. However, differences more become apparent when looking at geographical distance, population and GDP (Wu 2003).

For developing nations, geographically distance is a larger hurdle compared to more developed countries. It seems like these regions are mostly only covered by their neighbours while countries outside of their close proximity pay less attention to them.

In a sense, it seems like regionalism affects them more. However, aside from proximity, population size seems to be another significant predictor that is unique to developing nations. Perhaps poorer nations are considered to have a greater amount of importance and newsworthiness if they contains a larger part of the world population.

Proximity and population size have found to be less significant predictors when it comes to developed nations. For these countries, GDP appears to be the primary factor. As Wu mentions in his study, his findings could point to a trend of "treating international news as useful information about other countries". It is speculated that information about economically strong countries could be used to make better strategic decisions when it comes to negotiating and trading with them. This, combined with the high availability of news agencies in developed nations could further explain the reason behind the significance of these variables.

As a result, news flow often seems to be a one-way street. In their study, Himelboim, Chang, and McCreery (2010) found that developed nations generally stick to covering only a handful of other developed nations while developing countries do tend to cover developed nations more often. However, a more recent study suggests that the news flow ecology has evolved during the last decade (Guo and Vargo 2017). It seems that the historical dependence on US and western based news agencies is declining due to the emergence of more regional agencies and the rising prominence of the internet. Though it must be mentioned that while the landscape is gradually changing, news flow currently still seems to have a clear hierarchical nature.

If, as the literature suggests, it is possible to use machine learning to find distinct patterns between the news coverage styles of developing and developed nations, perhaps it would be possible to apply these techniques to predict or classify their distinct characteristics as well. For instance, perhaps shared patterns and biases can be found between nations who are experiencing similar levels of GDP growth, population growth or human development.

3. Methods

3.1 GDELT

The main data set that is used in this study is derived from GDELT Global Knowledge Graph 2.0 using an extraction algorithm written in Python. GDELT can be described as an extensive, ever updating data set that scrapes information from a plethora of news media sources from the internet every 15 minutes. After collecting the info, it extracts characteristics such as location references, subject matters (themes) and sentiment from each news item and encodes this so it can be used for big data analysis more effectively. While GDELT encodes information in many more categories, only the locations referenced in the item, the sentiment values and the sources will be used in this study. The news sources used by GDELT come from almost every country in the world and are automatically translated from 65 different languages (The GDELT Project 2015). This makes it a perfect data set to use for analysing patterns and biases in all corners of the globe instead of being restricted to the Anglosphere and other English sources. While the program does not natively include the country that a particular news source is affiliated with, an external data set was created by locating and mapping the relevant countries to the vast majority of sources that have been recorded by GDELT throughout the years (The GDELT Project 2018).

Unfortunately, GDELT 2.0 has less data available compared to version 1.0. The data created in this format only goes as far back as February 19, 2015 since it is a newer

version of the database. While this is a limitation, the current study is highly global in nature and the translation feature that is only present in version 2.0 was considered to be essential. Therefore the years that were included in the study range from 2016 up to 2019.

Due to hardware limitations, downloading every 15 minute interval of the 4 year period for both the translated and untranslated data was deemed impossible. Therefore it was decided to use 4 specific 15 minute intervals of each day instead. Since the data includes news items from nations all across the globe, time zones needed to be considered. This led to the decision to use the evenly spread out intervals of 00:00, 06:00, 12:00 and 18:00 to get a good coverage. Similarly, the use of a 4 year period was deemed important because, due to the high variation in yearly news content, patterns and biases will only show themselves consistently over a longer time frame. On rare occasions GDELT would not have data available for the preferred time slot on a specific day. The extraction algorithm therefore accounts for this by looking at each subsequent 15 minute interval until another available file presents itself. This strategy preserves the coverage of 4 distinct time slots per day while accounting for any possible missing data within the GDELT database.

3.2 Locations

As previously mentioned, GDELT extracts a wide array of variables from each news item and creates a different column for each of these. In the "Locations" column, each location that was mentioned in the item is stored using 7 different variables. These include the location type (Country, US state, US city, non-US state, non-US city), the full name of the location, the country code, the ADM1 code (2-character country code followed by another 2-digit code for the specific state or region), location latitude, location longitude and the location's feature ID (The GDELT Project 2015). While this allows for a great amount of specificity, the current study has an international focus and taking states and cities into account would have been out of the scope of this research. Therefore, each location was extracted by taking its country code, which uses the FIPS 140-2 format (Caddy 2005), and recording it in its own column. This led to a total of 263 countries that could potentially be referenced in each individual news item.

3.3 Sentiment

Additionally, sentiment values were also extracted from GDELT. There are a total of 7 variables that can be collected from GDELT's sentiment analysis. These include tone, positive score, negative score, activity reference density, self/group reference density and word count (The GDELT Project 2015). Tone consists of the average of positive and negative score. This fact makes the polarity variable important since this means that tone could potentially show a misleading neutral score in cases where both the positive and negative scores are high. Furthermore, activity reference density is a score that relates to the "activeness" of a text as opposed to a text that is more clinically descriptive. The self/group reference density variable captures the parts of a text that are self-referential or group-based in nature.

3.4 Other Data Sets

Aside from GDELT, multiple external data sets were used to capture data about the characteristics of each country. Several of these data sets were extracted from the World

Table 1
Sample of the country reference data set that was created using GDELT.

Source	Self	NL	RS	UK	US
Netherlands	0.396	0.018	0.027	0.079	0.109
Pakistan	0.508	0.004	0.067	0.110	0.217
Russia	0.523	0.015	0.055	0.071	0.098
Saudi Arabia	0.881	0.011	0.043	0.161	0.081
South Africa	0.474	0.000	0.019	0.091	0.151
United Kingdom	0.582	0.016	0.027	0.062	0.214

Bank. This is a global, financial organisation which, in their own words, "is one of the world's largest sources of funding and knowledge for developing countries". In addition, much of the data that is collected by them is open to the general public. The World Bank data used in this study includes data sets for GDP growth (World Bank 2019a) and population growth (World Bank 2019b). Additionally, the metadata for each of these data sets includes a classification for region and income group per country as defined by the World Bank.

Finally, a data set that includes a country's human development index (HDI) was used in this study. HDI is a score that reflects the human side of a country's developmental stage instead of a more economic approach such as GDP growth. The dimensions it takes into account when calculating the score are health, education and income per capita (UNDP 2019).

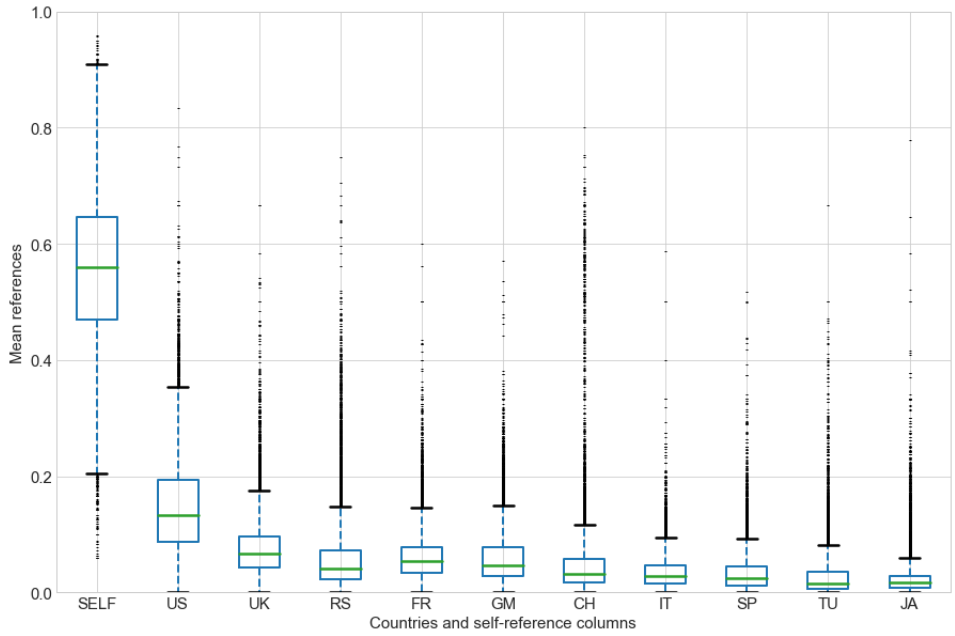
3.5 Preprocessing

Most of the columns that GDELT provides were irrelevant to the current study and could thus be dropped, leaving only the locations that were mentioned in the news item, its sentiment values and the nationality of its source. Both the country references and the sentiment data will be stored in their own separate data set where each row represents a week of a single country's news coverage data. For the country reference data set, each column represents a country that could potentially be referenced. The mean of these columns is taken for each time interval of a single day in order to create a matrix of 7 daily means per country. This process was repeated by taking the the mean of each week to construct a data set that contains 200 data points per country, each representing a weekly mean. Table 1 illustrates the composition of the final data set by showing a small sample of it. The sentiment data set was created using a similar method using the variables that were specified in the sentiment section as columns. In addition, all values were scaled to exist between 0 and 1 in order to make sure that the models could work more efficiently. These matrices represent the news coverage behaviours of a multitude of different countries over a 4 year period that can easily be analysed to explore any potential patterns and biases that may exist.

The countries that were used in the study were based on availability. Not every nation in the extracted data set had 4 full years worth of data. To make sure that the data was balanced, countries which did not meet these requirements were not included in the study. This left a total of 65 countries. To summarise, The full country reference

data set has a total of 13000 rows and 263 columns while the sentiment data set has 8 columns.

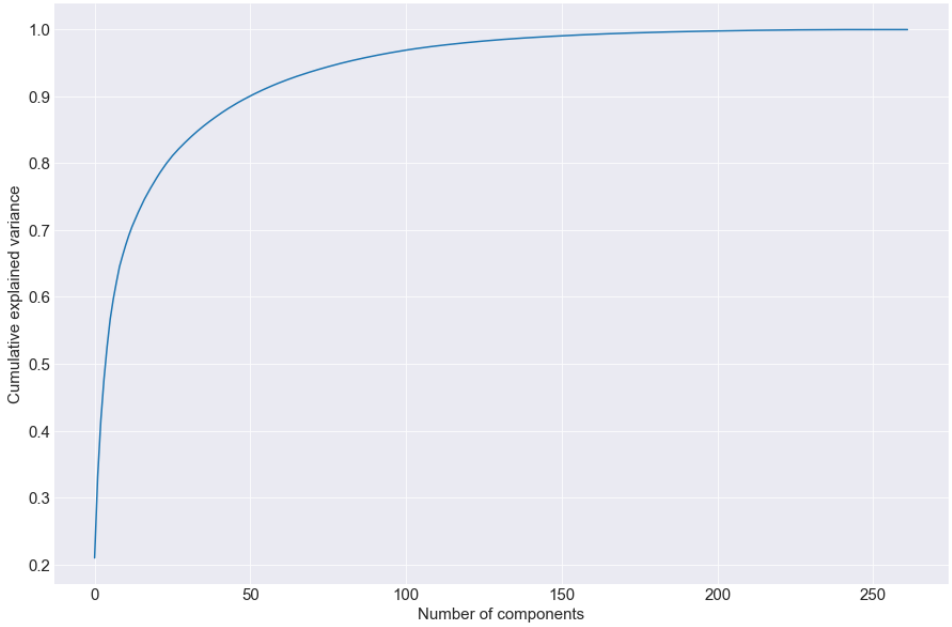
Figure 1
Boxplots of the self-reference column and the top 10 most referenced countries.



3.6 Normalisation of Self-References

Since it can be assumed that instances of domestic news coverage generally outnumber international coverage, an algorithm could potentially have an incredibly easy time classifying each country by observing which nation they cover the most. For example, whenever an unclassified source mentions the United States more often than any other nation, it is highly likely that this country is, indeed, the United States itself. This problem would have compromised the validity of the study. To deal with this issue, a specific "Self" column was created and any self-references were moved to this column instead. Of course, it would be unwise to leave an empty cell in place of the self-reference for similar reasons as have been stated previously so these particular cells were normalised by calculating the mean of the entire column to replace it. The value behind this method is in making sure that self-reference was not an outlier of any kind but rather an unremarkable number that would not sway the algorithm either way. By observing the boxplots of the self-reference column and the 10 countries with the highest mean references in figure 1, it can be observed that self-references are generally the most common type of mention, which supports the previous assumption. Additionally, it supports the literature in illustrating the same regions as the recipients of the most coverage.

Figure 2
PCA Cumulative variance against number of principal components.



3.7 Dimensionality Reduction

Since the predictors in this study include an array of 263 countries that could potentially be mentioned in a news item, the data is high dimensional in nature. Unfortunately, once the number of dimensions goes up, a data set becomes more sparse which in turn means that the required amount of data needed to make generalisations increases exponentially. To avoid the so-called curse of dimensionality, which reduces the effectiveness of machine learning models and increases the possibility of overfitting, Principal Component Analysis (PCA) has been used. PCA is an unsupervised machine learning technique that reduces the amount of dimensions while minimising the loss of information (Jolliffe and Cadima 2016). As can be seen in figure 2, if PCA transforms the data to have 50 components it will still retain 90% of the variance and up to 97% with 100 components. Therefore, it was decided to use 50 principal components in order to have a good balance between retaining the most amount of variance while minimising the dimensions to increase efficiency of the models and combat potential overfitting.

3.8 Experimental Setup

In this section, three distinct experiments will be discussed. Each of these experiments is designed to answer one of the three research questions and they will each be discussed in more detail below. A total of 200 data points per country will be used to analyse patterns and biases in news coverage behaviour that may emerge using several different methods. As previously discussed, each data point represents the mean value of a week's worth of news items and either the countries they have referenced or its sentiment values. Due to the exploratory nature of the research, the

different experiments will use a wide selection of both supervised and unsupervised machine learning techniques in order to get an expansive view of the subject and the effectiveness of the different approaches.

RQ1: To what extend can news coverage data and classification algorithms be used to identify countries?

For the first research question, the classification models will use the news source's nationality as primary target variable and the countries referenced in the article as predictors. News sentiment data will be extracted and used as an alternative predictor during the research as well to test its influence on the model's effectiveness and whether patterns can still be found while excluding regionalism as a factor.

To evaluate the classification model, the accuracy, precision, recall and F_1 score will be recorded. Since overfitting can be a large factor in deciding the effectiveness of a model, both the accuracy of the training data and that of the test data will be included for comparison. If the training accuracy is higher by a large margin, this could imply an overfitting issue. This will be evaluated against a baseline that represents random chance using the dummy classifier that is available from the Scikit-Learn library.

The entire data set, which ranges from 2016 up until 2019, will be used in this experiment. The data sets with country references, sentiment and the combination of the two are considered 3 different subsets.

RQ2: Which similarities between countries will become most apparent when clustering countries based on news coverage behaviour using unsupervised learning?

Since the ultimate goal of this question is to see which countries are clustered together, the data set in this experiment was compressed so that each data point is a single nation. This means that instead of having 200 data points per country, the mean per column of each country is calculated. Using hierarchical clustering methods, a dendrogram will be made which allows for precise visualisations of the process of clustering the high dimensional data. Indeed, transparency is key since the experiment is of an exploratory nature.

As an objective evaluation metric to give a concrete answer to the research question, homogeneity scores will be used. This score measures whether a cluster consists of similar values or many different ones. In order to add variables to measure cluster homogeneity on, the data set will be merged with information about each country's region and income group as classified by the World Bank (2019a). For developmental levels, an average HDI score for the years 2016 to 2018 will be added for each country as well. A high homogeneity in this case implies that each cluster is filled with countries from either the same region, the same economic class or with a similar average HDI. The optimal amount of clusters that the algorithm will eventually create will be decided by visually observing the dendrogram that is created in the previous phase of this experiment.

In order to better evaluate the value of the results, a baseline has been set up. Using a random number generator, an array of 7 random integers has been created to see how the other metric's homogeneity scores hold up against random chance. The reason for generating 7 integers is because both the region and HDI variables contain 7 different categories and it was designed to be similar in nature.

RQ3: Can supervised learning be used to explore news flow patterns that emerge from specific national characteristics instead of specific countries?

To test the hypothesis that there are distinct patterns that connect to national characteristics which can be predicted using the news data extracted from GDELT, regression algorithms will be applied on 3 different data sets. The sets included in this experiment are GDP growth, population growth and HDI, which will be used as target variables for the regression models. Since many nations did not have data for their HDI, GDP and population growth available for 2019 at the time of this experiment, only data from 2016 to 2018 was used.

Similarly to experiment 1, a dummy regression model will be used as a baseline performance. The strategy for the model in this case is to always suggest the mean of the data as output. The models will be evaluated on three different metrics: training and testing R^2 score to detect overfitting issues, and the mean squared error (MSE).

3.9 Software

To conduct the experiments, several scripts were written in Python 3.6 (Van Rossum and Drake 2009). Different scripts were coded to extract the data, analyse it, clean it of redundant or irrelevant variables, format it, preprocess it, apply machine learning algorithms to conduct the experiments and finally to visualise the results and processes.

The handling of the data was done by using both the pandas (McKinney 2010) and the NumPy (Van der Walt, S. and Colbert, S. C. and Varoquaux, G. 2011) libraries for Python. These tools were used to handle the data by transforming it into data frames and multidimensional arrays, which allows for quick and efficient wrangling of the data set. Furthermore, the tools used for preprocessing, hyperparameter tuning, clustering, classification and regression were obtained from the Scikit-learn library (Pedregosa et al. 2011). The visualisation libraries used in this study were Matplotlib (Hunter 2007) and Seaborn (Waskom et al. 2020).

3.10 Algorithms

It was was decided to use a selection of supervised and unsupervised learning algorithms in this study so it would be possible to to cross-reference the results and gain a deeper insight. Each of the models that were used will be discussed with a short explanation of their characteristics below. Firstly, the k-nearest neighbours algorithms will be discussed. Then the linear models, which include Naive Bayes and Ridge Regression. Subsequently, support-vector machines and the ensemble methods after which the clustering models will be touched upon. Each of these algorithms were implemented in python code using the Scikit-Learn library (Pedregosa et al. 2011).

Since it is often considered good practice to start as basic as possible, k-nearest neighbours (KNN) was chosen as the first model for its sheer simplicity. The algorithm works by comparing the current data point to a k number of data points that are closest in feature space. These are referred to as its neighbours. The classification variant compares it with the known classifications of its neighbours while it takes the average value of its neighbours for regression problems. KNN is a non-parametric algorithm which means that it has no prior assumptions about the data set. As a result, it is a very flexible and powerful model but it can be slow to run and is prone to overfitting (Muller and Guido 2017)).

Gaussian Naive Bayes represents a relatively simple, parametric classification algorithm. It learns its parameters by using Bayesian statistics on each of the different features. The naive part of the name refers to the fact that the algorithm assumes independence between variables. While in practice, this is often not the case, the model has a long history of performing well regardless.

The linear regression method that was used was Ridge regression, a variant of one of the simplest and oldest linear models for regression, ordinary least squares. This algorithm finds the parameters that minimise the mean squared error (MSE) between the prediction and true values. As an aside, MSE is also one of the metrics that will be used to evaluate the regression algorithms. While the basics of ridge regression are similar to regular linear regression, the addition of regularisation means that restrictions are added to the model to fight potential overfitting.

Support-vector machines appear in both a classification and regression form in this study. It is one of the most commonly used machine learning models and has a high versatility since it can approach problems in either a linear or non-linear fashion. It was chosen because of this flexibility and because it performs well in high dimensional and sparse data, which is very relevant in the mentioned countries data set. The support vector classifier will be referred to as SVC and the regression variant as SVR.

Random forests, an ensemble of decision trees, were included in the study. Decision trees use a series of binary decisions to form a tree-like structure to lead towards a final answer. A major drawback to this powerful tool is its sensitivity to overfitting. Random forests remedy this by creating a collection of them to cross-reference which gives the model a unique level of robustness. Furthermore, this algorithm allows for some much needed transparency in the form of feature importance analysis functionality. Additionally, to find the optimal hyperparameters for each algorithm, a grid search with cross validation was performed for every specific task.

For the clustering algorithm it was decided to use the hierarchical, agglomerative clustering model. Similarly to decision trees, it works in a hierarchical fashion that allows for transparency. The model begins by making each data point its own cluster and starts merging the ones that are closest to each other until there is only a single cluster left. This process can be visualised using a dendrogram and it has the option to stop clustering once it has reached a certain amount.

And finally, a dummy algorithm that makes random guesses as its classification strategy was used to provide a baseline for the others to beat. This is simply meant to indicate whether the other models perform better than chance or not.

In order to be certain that the algorithms would not be influenced by anything other than what was intended, a feature importance analysis was performed to to observe which columns of the data set were most influential in regards to performing machine learning tasks. Using a random forest algorithm, a model that utilises an ensemble of decision trees to classify the data, a list of the most important features was produced. These feature importance analyses are included in the results section of experiment 1.

4. Results

4.1 Experiment 1

In this first section, the classification performances on the country references data set will be evaluated. Subsequently the results of the feature importance analysis on this set will be described and the same process will be repeated on the sentiment data, and on a data set that combines the two. As can be seen in table 2, it appears to be

Table 2
Classification results from the country references data set.

Model	Train accuracy	Test accuracy	Precision	Recall	F_1 score
Baseline	0.02	0.02	0.02	0.02	0.02
KNN ($K = 4$)	0.87	0.81	0.82	0.81	0.80
Naive Bayes	0.80	0.77	0.79	0.77	0.77
Support Vector Classifier (C = 10)	**0.93**	**0.89**	**0.89**	**0.89**	**0.89**
Random Forest Classifier (max depth = 8)	0.84	0.77	0.79	0.77	0.75

very possible to classify countries based on which other nations they reference in their international news coverage. In fact, the classifiers all beat the baseline and produced very high results. While the baseline does not seem impressive, one must keep in mind that a data set with a high number of classes to classify makes it difficult to get any consistent results solely using random chance. Therefore, beating it indicates that the algorithms have actually identified patterns and biases in the data.

While all the classifiers have produced relatively similar results, SVC seems to perform best on this particular data set. A 89% test accuracy is a very impressive score. And while the training accuracy is slightly higher at 93%, a slight difference is to be expected due to the relatively small size of the data. It can be assumed that the gap will narrow once the amount of data is scaled up. The precision, recall and F1-score were

Figure 3
Feature importance analysis on country reference data.

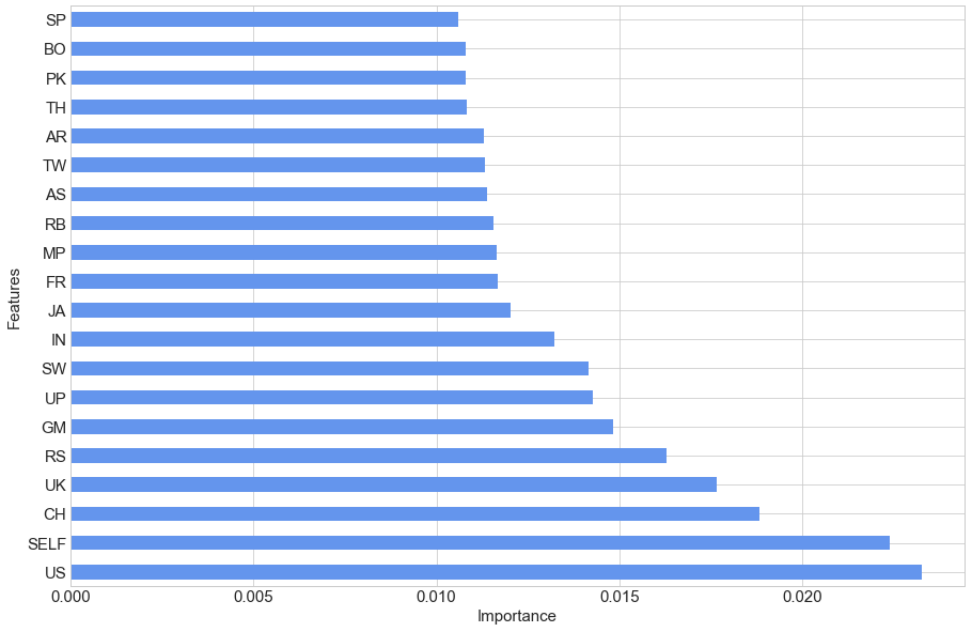


Table 3
Classification results from the sentiment data set.

Model	Train accuracy	Test accuracy	Precision	Recall	F_1 score
Baseline	0.02	0.02	0.02	0.02	0.02
KNN (n neighbours = 11)	0.61	**0.55**	0.52	**0.55**	**0.52**
Naive Bayes	0.53	0.52	0.52	0.54	0.51
Support Vector Classifier (C = 1)	**0.56**	**0.55**	**0.53**	**0.55**	**0.52**
Random Forest Classifier (max depth = 7)	0.57	0.53	0.53	0.53	0.49

all considered to be in order as well. To provide some more insight, a full classification report of the SVC model, in which the specific results for each specific country can be seen, is included in the appendix.

While both the KNN and random forest classifier seem perform admirably as well, the difference between the training and test accuracy is rather significant which suggest some slight overfitting issues. However, this was an expected limitation for both types of models.

Figure 3 depicts the results from the feature importance analysis. While the data set contains more features, it was decided to only display the top 20 here for readability purposes. A more comprehensive list can be found in the appendix.

Interestingly, both the US and the self-reference variable seem to be of more importance than the rest of the features. This makes sense considering these variables have a significantly larger amount of mentions and variability within their numbers compared to the other features as previously illustrated in figure 1. While many of the other high-ranking features could be explained by the fact that they are the most referenced countries, some potentially require a different explanation which will be elaborated on in the discussion section.

Despite the significantly less impressive results, table 3 illustrates that countries can indeed be classified using only sentiment data. This implies that countries still display distinct patterns from writing style and tone usage alone.

Again, SVC can be considered to give the best performance compared to the other algorithms. While both KNN and random forest produce similar results, their training scores indicate overfitting issues once again. SVC simply seems to suffer less from overfitting while still scoring relatively high. Naive Bayes, while reaching slightly lower heights than the other models, seems to be the most consistent in its results. This means that a case could be made for the idea that Naive Bayes is the most effective model in this situation.

When observing the feature importance analysis of this data in figure 4, it is interesting to note that each of the highest scoring features is related to writing style. Activity references, self/group references and word count seem to be the most influential while the sentiment values of average tone, negativity, positivity and polarity seem to be less significant in distinguishing countries.

The combination of the two data sets seems to yield a slight improvement in model performance over the country reference data set. Table 4 shows that SVC reaches 0.94 for test accuracy which is very impressive. This shows that, while sentiment alone does not produce the best results, its inclusion has significant value.

Table 4
Classification results from the combination of the country references and the sentiment data set.

Model	Train accuracy	Test accuracy	Precision	Recall	F_1 score
Baseline	0.02	0.02	0.02	0.02	0.02
KNN ($K = 4$)	0.91	0.88	0.88	0.88	0.87
Naive Bayes	0.84	0.84	0.86	0.84	0.84
Support Vector Classifier (C = 10)	**0.96**	**0.94**	**0.94**	**0.94**	**0.93**
Random Forest Classifier (max depth = 8)	0.91	0.85	0.86	0.85	0.84

4.2 Experiment 2

The results from the second experiment will be discussed in this section. Firstly, the dendrogram that visualises the clustering process will receive a short exploration after which it will be evaluated whether the clusters reflect similarities such as geographical, economic and developmental proximity

The dendrogram allows us insight in how the clustering algorithm makes its decisions. By observing this visualisation, it was decided that 17 clusters was an appropriate cutoff point for the rest of the experiment. This is visually displayed in the dendrogram in figure 5. While it can be seen that distinct regional groupings have formed in the data, some exceptions can not be explained this way and these will be reflected upon in the discussion section.

When observing the results from the homogeneity scores on the country reference data in table 5, it becomes obvious that geographical similarity is, indeed, the leading

Figure 4
Feature importance analysis on sentiment data.

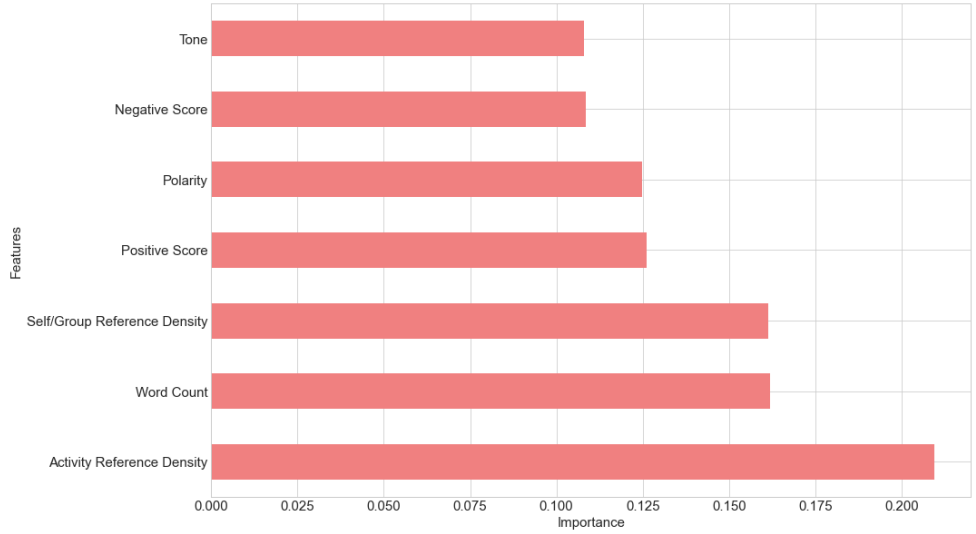
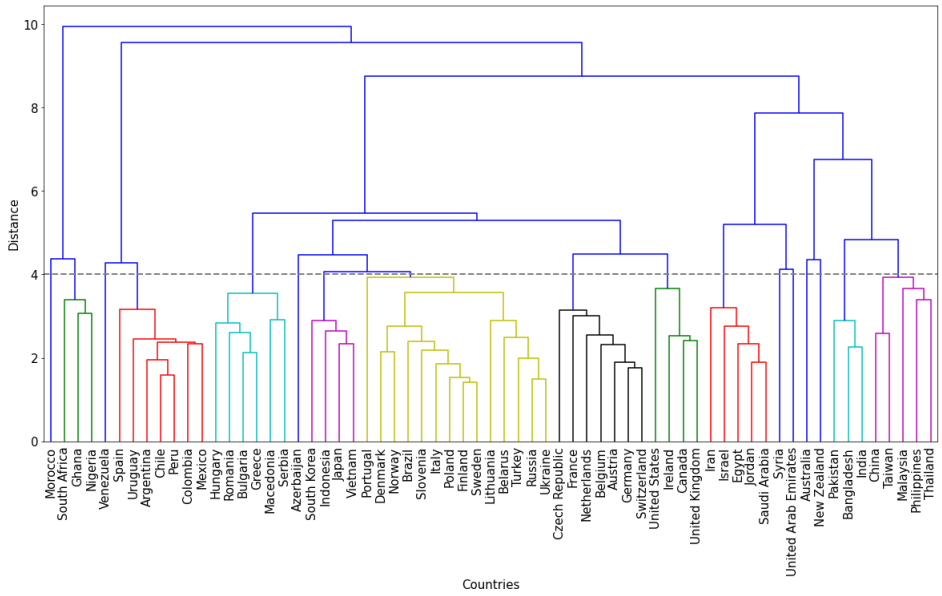


Figure 5
Hierarchical clustering dendrogram illustrating how each cluster is formed.



determinant when clustering. It must be mentioned that both income group and HDI seem quite homogeneous as well, though significantly less compared to region and only a 0.18 and 0.19 increase from the randomised baseline. While this is not dramatically higher, it is a significant difference. A full list of each country with their assigned cluster, region, income group and average HDI can be found in the appendix for reference.

Conversely, when clustering on sentiment data, the region homogeneity drops significantly. Unfortunately, both the income group and HDI homogeneity perform quite similarly to the random baseline and thus cannot be said to produce any meaningful results. In regards to the regional aspect of the sentiment based results, perhaps this could be another tangible connection between regional proximity and communication style when conveying news, even if the connection is not very strong. The combination of the data sets does not improve performance as much as with the classification task. In fact, region homogeneity seems to be negatively impacted while income group only increases by 0.02.

Table 5
Homogeneity scores for the clusters. This score reflects how homogeneous each cluster is.

	Baseline	Region	Income group	HDI
Country reference data	0.35	0.91	0.50	0.51
Sentiment data	0.37	0.54	0.36	0.37
Country reference & Sentiment data	0.31	0.83	0.52	0.51

4.3 Experiment 3

Table 6
Regression results on country reference data.

	Model	Train R^2 score	Test R^2 score	MSE
GDP growth	Baseline	-0.01	-0.01	6.35
	Ridge regression	0.18	0.18	4.76
	Random forest	0.40	0.32	4.27
	SVR	0.32	0.31	3.95
Population growth	Baseline	-0.02	-0.02	0.81
	Ridge regression	0.57	0.57	0.36
	Random forest	0.64	0.62	0.31
	SVR	0.78	0.76	0.19
HDI	Baseline	0.00	0.00	0.011
	Ridge regression	0.45	0.41	0.006
	Random forest	0.60	0.56	0.005
	SVR	0.69	0.64	0.004

In this final section, the performances of the regression models in regards to identifying GDP growth, population growth and HDI will be discussed. Good results in this case would suggest that countries who share similarities in regards to these characteristics, share distinct similarities within their news coverage behaviour as well.

Table 6 and 7 show that there are large differences in model performance between both the data sets and the regression tasks. On the country references data, it seems that the models perform well when it comes to population growth and HDI. SVR performs best on both the HDI and population growth tasks but does not produce equally good results with GDP growth. The same pattern appears with the sentiment data though the results are significantly less impressive. In addition, an experiment was run that

Table 7
Regression results on sentiment data.

	Model	Train R^2 score	Test R^2 score	MSE
GDP growth	Baseline	-0.01	-0.01	5.85
	Ridge regression	0.04	0.04	5.61
	Random forest	0.19	0.11	5.17
	SVR	0.10	0.10	5.33
Population growth	Baseline	0.00	0.00	0.81
	Ridge regression	0.03	0.02	0.79
	Random forest	0.19	0.16	0.67
	SVR	0.12	0.11	0.72
HDI	Baseline	0.00	0.00	0.013
	Ridge regression	0.15	0.15	0.009
	Random forest	0.31	0.26	0.008
	SVR	0.20	0.18	0.009

included a data set made from both the country reference and the sentiment data. The results of this experiment, however, had some slight improvements but were otherwise near identical to just the country reference data set. This might suggest that sentiment data has some value but not a large amount when it comes to this type of task. The table with the model performances on the combined data set can be found in the appendix.

5. Discussion

The original purpose of this study was to explore a novel machine learning-driven methodology in regards to studying patterns and biases that underlie international news coverage. To this end, many insights have presented themselves throughout the 3 different experiments.

Firstly, the results that the classification algorithms produced on the country reference data set clearly illustrate that each country in this study can automatically be identified by which other nations they choose to cover. Knowledge of these differences, of course, was already present in the literature. This experiment confirmed those findings in an automated manner and it shows that this method has merit. Additionally, the feature analysis performed produces some insights into the decision making process of these algorithms. In particular the apparent importance of countries such as Ukraine, Mauritius, Sweden, Argentina, Taiwan, Belarus, Spain and Pakistan. These nations are not prominently features in the 20 most covered countries. However, their importance could be explained by their relative obscurity. Since they would potentially only be mentioned in specific cases, they could prove to be essential in the classification process of these situations.

The second part of this experiment shows more than just a confirmation of what was already known. Since regionalism seemed to be the primary determinant in international news flow, the elimination of this variable illustrates that there may be more unexplored variables in this field that could be studied further. While the models do not perform as well, the fact that they could still produce consistent results and that combining the two data sets enhanced the results suggests that many countries have a distinct style in how they convey news that merits further analysis. What could the relation be between culture, language and the distinct way a country conveys news to its populace? What makes for these differences and what effect could it have on elements such as public opinion?

In the second experiment, the visualisation of the clustering process leads to the possibility of many interesting observations. For instance, the US, UK, Canada and Ireland have been grouped together even though they exist on two separate continents. Similarly, Spain has been grouped in with Latin America as well. Brazil is another notable exception, existing within a mostly European cluster, which could be caused by the language it shares with Portugal. These results appear to be quite consistent with the literature (Kim and Barnett 1996). The consistency of these results after many years and the emergence of the internet, strengthens the idea that similarities in language, culture and historical colonial ties are an important determinants in news flow. Notable differences between the findings can be found in the clearly defined Middle-Eastern and African clusters. This suggests a potential increase in the independence of these regions since the previous study was conducted and perhaps a lower reliance on news agencies from more developed nations and could support the findings of Guo and Vargo (2017).

Exceptions aside, region seems to be the primary factor behind the formation of these clusters. Cluster homogeneity was apparent with economic and developmental factors as well but with lower results. Unfortunately, the results from the sentiment

data set were mostly insignificant for the latter two variables. It must be mentioned that a limitation in this regard was the limited data about countries with either a low income group or HDI. While they did appear in the GDELT data with some frequency, it was rare for such a country to have weekly averages for the entire duration of the study. However, their inclusion would have made the data set very imbalanced and they were therefore left out if there was not sufficient data available. Perhaps a more complete data set would have painted a more complete picture as well.

Just like in the previous experiment, however, there was still a link present between sentiment data and region. Based on the literature, a data set about international trade could potentially lead to interesting results in this experiment as well. Unfortunately, no suitable data set could be found for use in this study. Any analysis that could be done with the available, complex trade data would deserve a thesis of its own and was therefore out of the scope of this study. Future research could focus on using this methodology to explore trade and news flow patterns. The literature does link trade to news flow and improving the understanding of the underlying structures of international trade can prove to be useful for many people.

Finally, the regression experiments. The primary goal was to investigate whether news coverage behaviour was tied to national characteristics as well as countries. For instance, do countries that are undergoing a big GDP growth that year share some patterns and biases? The results seem to suggest that, indeed, certain patterns emerge in tandem with certain characteristics. Due to the exploratory nature and breadth of the current study, there is plenty of unexplored depth to this particular aspect. While this is a limitation, it shows a lot of potential for future research.

Since it appears that patterns in news flow provide some predictive value for national characteristics, perhaps there is much more can be done with this subject. Potentially, future studies could utilise deep learning techniques such as recurrent neural networks to forecast GDP or population growth for many years in advance. Such a study could provide a new methodology and perspective to this area of research.

6. Conclusion

Can supervised and unsupervised machine learning techniques be used to effectively explore patterns and biases in international news flow?

This overarching research question is what led to the breadth of techniques that were employed to explore a wide variety of aspects within this topic. The many insights about the underlying structures of news flow lead to the conclusion that there is a lot of potential within this approach. This study's use of machine learning techniques could be expanded upon by applying it to alternate data sets, as this study has done with sentiment. GDELT provides access to references of organisations or themes discussed per article which could lead to new and interesting findings as well. For instance, the current study's findings in regards to sentiment could spark interesting questions about the relationship between communication style and culture, region or language.

As previously stated, it is believed that using the knowledge of news data patterns and national characteristics especially, holds a large amount of potential. The potential ability to forecast a variety of variables on a global scale using a country reference data set, which is easily available through GDELT, has huge implications for future research. In a globalised society, the importance of understanding international news flow and studying the potential applications of this knowledge can not be understated.

7. Self-Reflection

Working on this thesis project was a very valuable learning experience in several different aspects. Firstly, I have significantly deepened my knowledge about data science and the techniques that I have been studying for the past few years. The exploratory nature and the breadth of the project allowed me to apply a variety of both supervised and unsupervised models and subsequently, gain new insights about their practical applications. For instance, visualising the clustering model was a challenge at first. It required me truly understand the underlying mechanics before I could overcome it and visualise it in an appropriate manner. This also includes learning a lot about how to handle large amounts of data and how to create a usable and effective data set out of raw data.

Moreover, since this project had very specific requirements in regards to the data, which the GDELT library could not satisfy, my coding skills have improved significantly. Finding creative solutions to abstract problems and implementing this successfully using Python was a large part of my learning process. Problems that would have taken me days at the start could be solved in a fraction of that time near the end of the project.

Finally, I believe that I have learned a lot about what it means to conduct research. This includes improving the important skills that are associated with this such as academic writing and time management. I had a lot of difficulty with finding a subject and committing to it. However, as time passed and the project evolved, I learned the importance of literature and reflecting upon research that was already done in order to find a way forward. For future projects, I will therefore start earlier with intensively reviewing the literature to avoid a lot of the issues I faced early on.

References

Aalberg, T., S. Papathanassopoulos, S. Soroka, J. Curran, K. Hayashi, S. Iyengar, P. K. Jones, G. Mazzoleni, H. Rojas, D. Rowe, and R. Tiffen. 2013. International tv news, foreign affairs interest and public knowledge. *Journalism Studies*, 14(3):387–406.

Caddy, T. 2005. *FIPS 140-2*. Springer US, Boston, MA.

Chang, T. K., P. J. Shoemaker, and N. Brendlinger. 1987. Determinants of international news coverage in the u.s. media. *Communication Research*, 14(4):396–414.

Guo, L. and C. J. Vargo. 2017. Global Intermedia Agenda Setting: A Big Data Analysis of International News Flow. *Journal of Communication*, 67(4):499–520.

Himelboim, I., T. K. Chang, and S. McCreery. 2010. International network of foreign news coverage: Old global hierarchies in a new online world. *Journalism & Mass Communication Quarterly*, 87(2):297–314.

Hunter, J. D. 2007. Matplotlib: A 2d graphics environment. *Computing in Science Engineering*, 9(3):90–95.

Jolliffe, I. T. and J. Cadima. 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

Kariel, H. G. and L. A. Rosenvall. 1984. Factors influencing international news flow. *Journalism Quarterly*, 61(3):509–666. Retrieved from: https://journals.sagepub.com/doi/pdf/10.1177/107769908406100305?casa$_t oken = XUmtzvQHQxcAAAAA : PG_s DFgOMEW - Z08GF4ZjZmZpoTXNAdomUD - jQy - UlUspqHgifWwpxybLTadA0yH_p DDRgPCa6Trz6w$.

Kim, K. and G. A. Barnett. 1996. The determinants of international news flow: A network analysis. *Communication Research*, 23(3):323–352.

Kwak, H. and J. An. 2014. Understanding news geography and major determinants of global news coverage of disasters. Retrieved from: https://arxiv.org/abs/1410.3710.

McKinney, W. 2010. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, pages 51 – 56.

Muller, A. C. and S. Guido. 2017. *Introduction to machine learning with Python: A guide for data scientists.* O'Reilly Media, New York.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Segev, E. 2015. Visible and invisible countries: News flow theory revised. *Journalism*, 16(3):412–428.

The GDELT Project. 2015. Gdelt 2.0 global knowledge graph codebook (v2.1). Retrieved from: https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/.

The GDELT Project. 2018. *Mapping the Media: A Geographic Lookup of GDELT's Sources.* Retrieved from: https://blog.gdeltproject.org/mapping-the-media-a-geographic-lookup-of-gdelts-sources.

UNDP. 2019. Human development data (1990-2018). Retrieved from http://hdr.undp.org/en/data.

Van der Walt, S. and Colbert, S. C. and Varoquaux, G. 2011. The numpy array: A structure for efficient numerical computation. *Computing in Science Engineering*, 13(2):22–30.

Van Rossum, G. and F. L. Drake. 2009. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

Wanta, W., G. Golan, and C. Lee. 2004. Agenda setting and international news: Media influence on public perceptions of foreign nations. *Journalism & Mass Communication Quarterly*, 81(2):364–377.

Waskom, M., O. Botvinnik, J. Ostblom, M. Gelbart, S. Lukauskas, P. Hobson, D. C. Gemperline, T. Augspurger, Y. Halchenko, J. B. Cole, J. Warmenhoven, J. De Ruiter, C. Pye, S. Hoyer, J. Vanderplas, Villalba. S., G. Kunter, E. Quintero, P. Bachant, M. Martin, K. Meyer, C. Swain, A. Miles, T. Brunner, D. O'Kane, T. Yarkoni, M. L. Williams, C. Evans, C. Fitzgerald, and Brian. 2020. mwaskom/seaborn: v0.10.0 (january 2020).

Wilke, J., C. Heimprecht, and A. Cohen. 2012. The geography of foreign news on television: A comparative study of 17 countries. *International Communication Gazette*, 74(4):301–322.

World Bank. 2019a. Gdp growth (annual %). Retrieved from: https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG.

World Bank. 2019b. Population growth (annual %). Retrieved from: https://data.worldbank.org/indicator/SP.POP.GROW.
Wu, H. D. 2000. Systemic determinants of international news coverage: a comparison of 38 countries. *Journal of Communication*, 50(2):110–130.
Wu, H. D. 2003. Homogeneity around the world?: Comparing the systemic determinants of international news flow between developed and developing countries. *Gazette (Leiden, Netherlands)*, 65(1):9–24.

Appendix

Table 8
Regression results on country reference & sentiment data.

	Model	Train R^2 score	Test R^2 score	MSE
GDP growth	Baseline	-0.01	0.00	5.82
	Ridge regression	0.21	0.19	5.82
	Random forest	0.41	0.29	4.11
	SVR	0.32	0.31	4.01
Population growth	Baseline	-0.00	-0.00	0.75
	Ridge regression	0.59	0.54	0.31
	Random forest	0.65	0.59	0.33
	SVR	0.78	0.75	0.20
HDI	Baseline	0.00	0.01	0.110
	Ridge regression	0.50	0.47	0.006
	Random forest	0.63	0.59	0.005
	SVR	0.68	0.65	0.004

Table 9
Full SVC classification report on the country reference data.

Country	Precision	Recall	F_1 score	Support
Argentina	0.87	0.94	0.91	51
Australia	0.82	0.82	0.82	44
Austria	0.90	1.00	0.95	44
Azerbaijan	0.98	0.96	0.97	53
Bangladesh	0.85	0.83	0.84	48
Belarus	1.00	0.96	0.98	45
Belgium	0.85	0.98	0.91	42
Brazil	0.82	0.98	0.90	48
Bulgaria	0.87	0.96	0.91	48
Canada	0.83	0.98	0.90	66
Chile	0.77	0.66	0.71	56
China	1.00	1.00	1.00	46
Colombia	0.88	0.94	0.91	49
Czech Republic	0.93	0.86	0.90	50
Denmark	0.78	0.60	0.68	53
Egypt	0.84	0.91	0.88	54
Finland	0.78	0.85	0.82	47
France	0.92	0.98	0.95	58
Germany	0.85	0.94	0.89	47
Ghana	0.77	0.72	0.74	46
Greece	0.98	0.98	0.98	46
Hungary	0.85	1.00	0.92	44
India	0.95	1.00	0.98	40
Indonesia	0.89	1.00	0.94	50
Iran	0.93	0.90	0.91	48
Ireland	0.80	0.78	0.79	55
Israel	1.00	1.00	1.00	57
Italy	0.90	1.00	0.95	47
Japan	0.93	1.00	0.96	39
Jordan	0.98	0.91	0.94	46
Lithuania	0.95	0.82	0.88	49
Macedonia	0.96	0.85	0.90	53
Malaysia	0.82	0.71	0.76	52
Mexico	0.92	0.98	0.95	57
Morocco	0.89	0.78	0.83	50
Netherlands	0.86	0.86	0.86	51
New Zealand	0.89	0.78	0.83	50
Nigeria	0.92	0.90	0.91	61
Norway	0.74	0.62	0.67	47
Pakistan	0.94	0.94	0.94	50
Peru	0.69	0.78	0.73	45
Philippines	0.87	0.82	0.85	50
Poland	0.86	0.92	0.89	53
Portugal	0.88	0.92	0.90	53
Romania	0.91	0.83	0.87	48
Russia	1.00	1.00	1.00	50
Saudi Arabia	0.90	0.87	0.89	54
Serbia	1.00	1.00	1.00	52
Slovenia	0.93	0.76	0.83	49
South Africa	0.69	0.63	0.66	46
South Korea	0.91	0.91	0.91	43
Spain	0.88	0.98	0.93	45
Sweden	0.84	0.84	0.84	56
Switzerland	0.94	0.98	0.96	49
Syria	0.94	0.98	0.96	50
Taiwan	0.96	0.98	0.97	50
Thailand	0.94	0.92	0.93	49
Turkey	0.95	0.96	0.95	54
Ukraine	0.95	0.98	0.97	62
United Arab Emirates	0.85	0.79	0.81	42
United Kingdom	0.87	0.95	0.91	43
United States	0.91	0.98	0.94	50
Uruguay	0.87	0.50	0.63	52
Venezuela	0.95	0.95	0.95	61
Vietnam	0.95	1.00	0.97	57
accuracy			0.89	3250
macro avg	0.89	0.89	0.89	3250
weighted avg	0.89	0.89	0.89	3250

Figure 6
Feature importance analysis on country reference data. Top 100 features

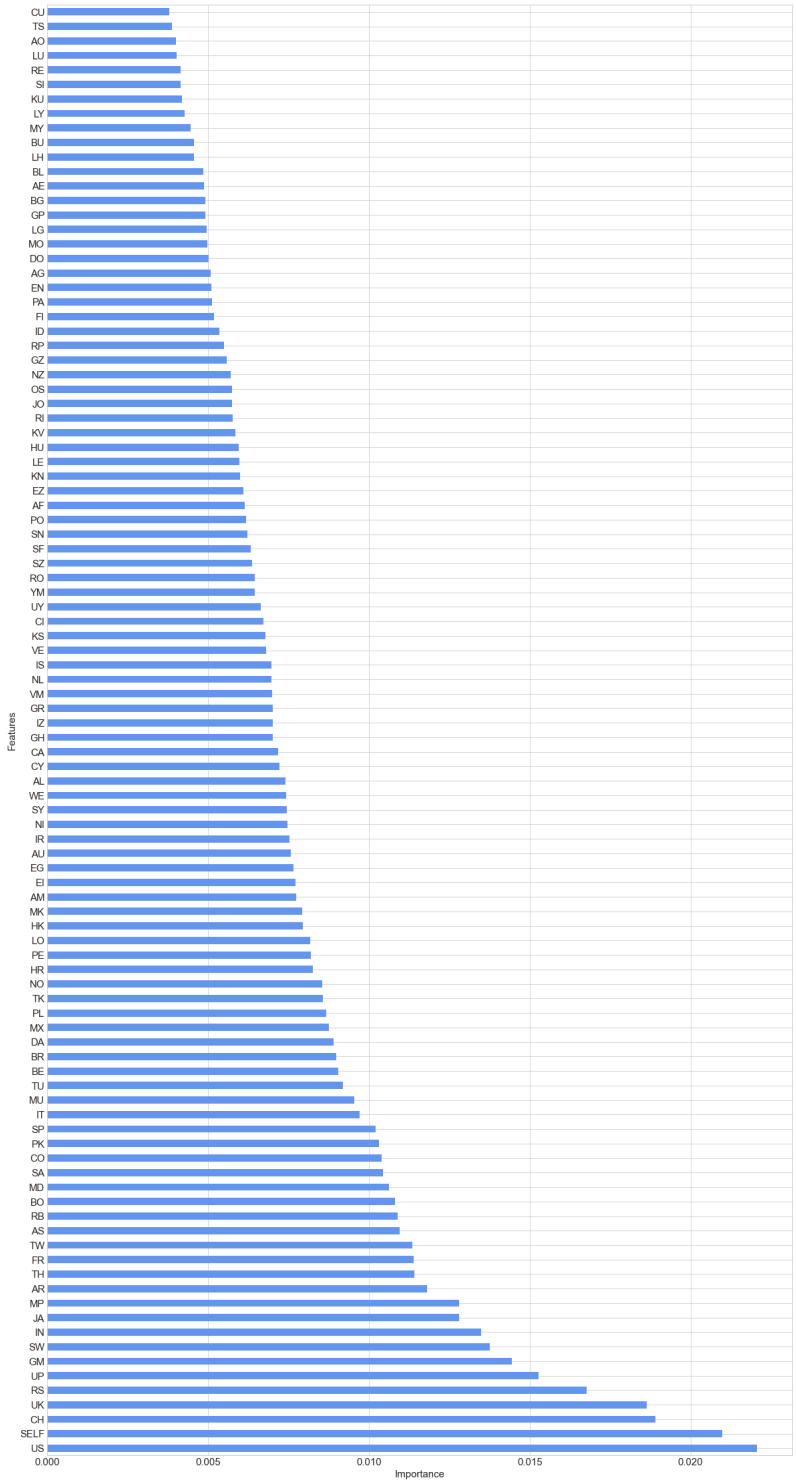


Table 10
Full cluster results.

Cluster	Country	Region	IncomeGroup	HDI Score
0	Finland	Europe & Central Asia	High income	9.0
0	Lithuania	Europe & Central Asia	High income	8.0
0	Norway	Europe & Central Asia	High income	9.0
0	Italy	Europe & Central Asia	High income	8.0
0	Ukraine	Europe & Central Asia	Lower middle income	7.0
0	Belarus	Europe & Central Asia	Upper middle income	8.0
0	Turkey	Europe & Central Asia	Upper middle income	8.0
0	Brazil	Latin America & Caribbean	Upper middle income	7.0
0	Poland	Europe & Central Asia	High income	8.0
0	Portugal	Europe & Central Asia	High income	8.0
0	Russia	Europe & Central Asia	Upper middle income	8.0
0	Slovenia	Europe & Central Asia	High income	8.0
0	Sweden	Europe & Central Asia	High income	9.0
0	Denmark	Europe & Central Asia	High income	9.0
1	Philippines	East Asia & Pacific	Lower middle income	7.0
1	Malaysia	East Asia & Pacific	Upper middle income	8.0
1	Taiwan	East Asia & Pacific	High income	9.0
1	China	East Asia & Pacific	Upper middle income	7.0
1	Thailand	East Asia & Pacific	Upper middle income	7.0
2	Macedonia	Europe & Central Asia	Upper middle income	7.0
2	Serbia	Europe & Central Asia	Upper middle income	7.0
2	Greece	Europe & Central Asia	High income	8.0
2	Hungary	Europe & Central Asia	High income	8.0
2	Romania	Europe & Central Asia	Upper middle income	8.0
2	Bulgaria	Europe & Central Asia	Upper middle income	8.0
3	United States	North America	High income	9.0
3	United Kingdom	Europe & Central Asia	High income	9.0
3	Canada	North America	High income	9.0
3	Ireland	Europe & Central Asia	High income	9.0
4	Peru	Latin America & Caribbean	Upper middle income	7.0
4	Spain	Europe & Central Asia	High income	8.0
4	Uruguay	Latin America & Caribbean	High income	8.0
4	Mexico	Latin America & Caribbean	Upper middle income	7.0
4	Argentina	Latin America & Caribbean	Upper middle income	8.0
4	Chile	Latin America & Caribbean	High income	8.0
4	Colombia	Latin America & Caribbean	Upper middle income	7.0
5	South Africa	Sub-Saharan Africa	Upper middle income	7.0
5	Nigeria	Sub-Saharan Africa	Lower middle income	5.0
5	Ghana	Sub-Saharan Africa	Lower middle income	5.0
6	Venezuela	Latin America & Caribbean	Upper middle income	7.0
7	South Korea	East Asia & Pacific	High income	9.0
7	Indonesia	East Asia & Pacific	Lower middle income	7.0
7	Vietnam	East Asia & Pacific	Lower middle income	6.0
7	Japan	East Asia & Pacific	High income	9.0
8	Israel	Middle East & North Africa	High income	9.0
8	Saudi Arabia	Middle East & North Africa	High income	8.0
8	Jordan	Middle East & North Africa	Upper middle income	7.0
8	Egypt	Middle East & North Africa	Lower middle income	6.0
8	Iran	Middle East & North Africa	Upper middle income	7.0
9	India	South Asia	Lower middle income	6.0
9	Bangladesh	South Asia	Lower middle income	6.0
9	Pakistan	South Asia	Lower middle income	5.0
10	Austria	Europe & Central Asia	High income	9.0
10	Belgium	Europe & Central Asia	High income	9.0
10	Netherlands	Europe & Central Asia	High income	9.0
10	Switzerland	Europe & Central Asia	High income	9.0
10	Czech Republic	Europe & Central Asia	High income	8.0
10	France	Europe & Central Asia	High income	8.0
10	Germany	Europe & Central Asia	High income	9.0
11	Azerbaijan	Europe & Central Asia	Upper middle income	7.0
12	Morocco	Middle East & North Africa	Lower middle income	6.0
13	Australia	East Asia & Pacific	High income	9.0
14	New Zealand	East Asia & Pacific	High income	9.0
15	Syria	Middle East & North Africa	Low income	5.0
16	United Arab Emirates	Middle East & North Africa	High income	8.0