Lappeenranta University of Technology
School of Business and Management
Degree Program in Computer Science

Bachelor’s thesis

Varpu Huhtinen

Analyzing how viewpoint regarding climate change has evolved in news stories since 1989 by applying topic modelling

Work supervisor: Post-doctoral researcher Annika Wolff

TIIVISTELMÄ

Lappeenrannan teknillinen yliopisto
School of Business and Management
Tietotekniikan koulutusohjelma

Varpu Huhtinen

Miten uutisoiminen ilmastonmuutoksesta on muuttunut vuodesta 1989 lähtien aihemallinnus ohjelman mukaan.

Kandidaatintyö

2019

33 sivua,1 taulukko

Työn tarkastajat: Tutkijatohtori Annika Wolff, Apulaisprofessori Jussi Kasurinen

Hakusanat: akateeminen opinnäytetyö, Ilmastonlämpeneminen, piilevä Dirichlet-jako, koneoppiminen, aihemallinus
Keywords: academic thesis, Climate Change, Latent Dirichlet Allocation, machine learning, topic modelling

Mielipiteet ilmastonmuutosta kohtaan ovat muuttuneet huomattavasti vuosien varrella ja kansan suurimman osan tärkein tietolähde on uutislähteet. Tämän takia, työssä tutkitaan piilevää Dirichlet-jakoa hyödyntäen, miten uutisoiminen ilmastonmuutoksesta on muuttunut vuodesta 1989 lähtien. Tuloksina huomattiin, miten epävarmuus ilmastonlämpenemistä kohtaan on kehittynyt ja milloin tutkimusta on alettu kehittämään. Tuloksista myös ilmeni keskeisimmät aiheet ilmastonlämpenemisessä eri vuosiluvuilla. Työ toimii taustana kattavammalle aihelmallinnus tutkimukselle pienen uutislähde määrän takia.

ABSTRACT

Lappeenranta University of Technology
School of Business and Management
Degree Program in Computer Science

Varpu Huhtinen

Analyzing how viewpoint regarding climate change has evolved in news stories since 1989 by applying topic modelling

Bachelor's Thesis

33 pages, 1 table

Examiners: Postdoctoral researcher Annika Wolff, Assoc. Prof. Jussi Kasurinen
Keywords: academic thesis, Climate Change, Latent Dirichlet Allocation, machine learning, topic modelling

The viewpoint towards climate change has changed in many ways throughout the years and the main information source for people is newspapers. In this study we use Latent Dirichlet Allocation to find how the viewpoint in news stories towards climate change has evolved since 1989. The results highlighted how the incertitude towards climate change has evolved and when the studies have been started. In the study, we also found out the most important topics related to climate change. This study can be used as a base for a larger topic modelling study on news stories associated with climate change due to the small amount of newspapers used as corpora.

TABLE OF CONTENT

LIST OF SYMBOLS AND ABBREVIATIONS

API	Application Programming Interface
IPCC	Intergovernmental Panel on Climate Change
LDA	Latent Dirichlet Allocation
MALLET	Machine Learning for LanguagE Toolkit

1 INTRODUCTION

1.1 Background

We live today in a world where technologies, vehicles and modern consuming manners are part of our everyday routine, but all these things that we are used to do or to use are influencing the climate change that is taking place right now. One of the biggest reasons for climate change is the release of greenhouse effect gases into the atmosphere and these gazes are undeniably caused by our modern society's living habits. In our everyday life we use our car to go to work, we get onto planes to go to business meetings or to go on holidays, and all this transports are huge greenhouse gas emitters (Smith and Rodger, 2009; Stanley et al., 2011). Even our dietary choices are a huge greenhouse gas effect source mainly if we are used to eat meat or rice (Carlsson-Kanyama, 1998) .

According to the Atlas of Climate Change (Dow and Downing, Thomas E, 2006), the transport sector has emitted a significant amount of greenhouse effect gases into the atmosphere since 1990, as in 2000 the emission had increased of 36% compared to 1990. In 2003, 70% of the gas emitted by transportation sector were from cars and trucks, which means by our everyday activity as trucks are used for the transportation of food and goods that we use daily.

Climate change affects each living being on earth which means humans, animals, insects and even plants. The impacts can go from the changing in Arctic sea ice to changes in coral reef ecosystems. During the past years the amount of natural disasters such as hurricanes, floods and wildfires have increased widely, and it is mainly due to weather changes. Other changes notable in the environment are the glacial retreats in mountains which glaciers are losing volume or the animal's behavior which is changing as well. Some birds or butterflies have indeed changed their migration habits due to the climate change.

The importance of awakening the consciences is now more important than ever as, according to a recent report by Intergovernmental Panel on Climate Change (IPCC), the global warming might reach 1.5 °C in 2052 if nothing is done to reduce its increasing speed.("IPCC presents findings of the Special Report on Global Warming of 1.5°C at event to discuss Viet Nam's response to climate change — IPCC," 2018)

Even though the climate change today is undeniable, during the past thirty years it has been a controversial subject and it impacts have not always been clear. Many have doubted or are doubting its reality and even today the real consequences are questioned by some people.

Analyzing the news stories on this subject throughout the past thirty years can give a good idea of how the opinion on climate change has been affected during this time-span. And we will be able to see how medias have been relaying the information about this severe problem.

Analyzing how news stories relate the news linked to climate change is indeed important, as news stories are one of the main ways to give information to a major part of the population. Analyzing them will give a good idea of what is and has been the most important topic during the past years and how the topic importance has changed year after year.

1.2 Goals and limits

This study's goal is to find how the public opinion has changed over the past thirty years regarding climate change by defining multiple topics using topic modelling. The principle goals are to find when the climate change problem really became a big issue and when finding solutions against it became notable. It is also interesting to find how the skepticism towards climate change consequences has evolved during the studied time-span. Moreover, it is interesting to analyze if specific events increased the covering of climate change by media.

As this study is mainly focusing on climate change and the opinions related to it, new stories about the use of renewable energies are not part of the study. Neither are the solutions employed to reduce the greenhouse effect or the opinions related to them. The new stories selected as a corpus for this study will all be in English as the LDA is working with only one language at a time.

This study is relying on about fifty new stories from 1989 to today related to climate change. Topic modeling, which is a type of statistical modelling, is used to find the main ideas of these articles. The precise name of the topic modelling type used is Latent Dirichlet Allocation (LDA).

1.3 Work structure

In the first part of this study, we set the background of the work: why it is done and what will be done. The limits and goals of the work were also defined.

In the second part, first we define what LDA is and then there is a literature overview of some topic modelling done before using LDA as well as studies about public's concern on climate change and an analyze how these literatures can be re-used in the context of this study.

In the result section the LDA program is presented and explained as well as the result given as an output by the program. Along with the result table, an analyzes timespan by timespan is presented of the topics found by LDA.

In the discussion section the way the program affects the result is discussed. A parallel with studies and events of each timespan is also made.

Then, there is the reflection and future part where we define how the study could be improved and made more precise.

2 THESIS

2.1 Latent Dirichlet Allocation (LDA)

According to Jacobi et al. topic modelling allows us to analyze large amount of data without manual coding and LDA allows us to code new stories automatically to create topic based on the document's patterns. When using LDA, the output of the program is a set of clusters of words and each of them have internal consistency which meaning must be found and interpreted by the searcher in order to find the response to the study. Moreover, LDA considers every word of a document which means that even unpopular topics are also considered. As unpopular topics are considered, it leaves the searcher to filter the not relevant clusters out of the study. If emotions are widely expressed in news articles, the cluster could also represent the writing style of the writer. (Jacobi et al., 2016)

In order to use LDA we need a corpus M of documents such as $D = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_M\}$ where $\boldsymbol{w} = (w_1, w_2, \ldots, w_N)$ is the document composed of a sequence of N words such as w_N is the *n*th word in the sequence. Each $\boldsymbol{w}_i$ in D, is supposed by LDA to be generated following the process below:

1. Choose $N \sim Poisson(\xi)$
2. Choose $\theta \sim Dir(\alpha)$
3. For each w_N:

   (a) Choose a topic $z_N \sim Multinomial(\theta)$

   (b) Choose a word w_N from the multinomial probability $p(w_N|z_N, \beta)$ conditioned on the topic z_N

Where θ is a mixture over *k* topics in each document defined by the *k*-dimensional parameter of a Dirichlet distribution α. A topic z_N is assigned by sampling from θ to every N word position in the document. β is the variable which defines a multinomial distribution over the vocabulary of each topic (z_N) when a word w_N is selected from $p(w_N|z_N, \beta)$ once the word position's is known. (Blei et al., 2003)

To summarize, LDA is a generative model used to find distinct topics in a corpus of documents. It assumes that every document is a mixture of topics and that each topic is a mixture of words. The goal of LDA is to find the most relevant topics for a corpus of document as well as the most relevant words for a specific topic. (Bolelli et al., 2009)

2.2 Literature overview

2.2.1 Application of LDA

Since it was first introduced in 2003 by Blei et al. LDA has been used in many purposes and extensions have been added to allow it to work on more various documents. One of the most notable extension over the years has been to make LDA usable in the computer vision field. It has been used to identify objects and to assign them into categories by replacing words by visual analogues of them formed by vectors (Sivic et al., 2005) and even to categorize not only objects but entire scenes (Fei-Fei and Perona, 2005). LDA has also been a security provider as it can be applied to Web spam filtering (Bíró et al., 2008) or even to detect telecommunication fraud (Xing and Girolami, 2007). One other application context of using LDA is forecasting events such as responses to political posts (Yano et al., 2009) or popularity of social media publications (Hong et al., 2011). In this paper, even though LDA can be applied to several document types, we will focus only on applying it to news stories. Therefore, below are detailed two application examples of LDA on news stories which are somehow closed to the subject of this study and can consequently be used to improve this work.

2.2.1.1 Analyzing nuclear technology articles with LDA

In 2016 in their study about analysis of large amounts of journalistic texts using topic modeling, Arina Jacobi et al. Applied LDA on new stories about nuclear technology and compared the result with the study done by Gamson and Modigliani in 1989, where they studied the culture surrounding nuclear technology.

The goal of the study led by Arina Jacobi et al. Was to find if the subject of nuclear technology was discussed in the same way from 1945 to 2013. The study was done by selecting news articles in New York Times Application Programming Interface (API) by performing a search with the word “nuclear”.

The topics found by LDA seemed more specific and concrete than the frames found by Gamson and Modigliani in their study. However, both studies found that the topics evolution did change over time, but they did not increase or decrease in a linear way. The viewpoints associated with nuclear technology were not found as clearly by LDA than they were by Gamson and Modigliani as in their study they did not use only words but also images and videos. One of the main topics that emerged from the LDA study was still the concern of the threat of nuclear energy. The topics found by LDA were not all relevant and they affected the quality of the output, so it was found that filtering useless information before performing the LDA would give more precise and relevant topics. Useless information are places where the articles are written, book or film reviews, index articles or news summaries. However, as these irrelevant topics were clustered together it was also possible to take them of the study afterwards. (Jacobi et al., 2016)

In the study covered in this document we are looking for the evolution of the viewpoints associated with climate change which follows the same idea as the study done by Jacobi et al. It is important to note that finding clear viewpoint is not a trivial process and that some information present in the news articles is not relevant. To avoid not relevant information, it is important to find news articles that cover only climate change and not anything else such as reviews of documents about climate change. As the corpus is done manually, leaving irrelevant new stories from it is an easy task.

2.2.1.2 Applying LDA to find relevant topics in historical newspaper

In 2011 Yang et al. led a study about applying topic modelling on historical newspaper with the goal to identify the most important and interesting topics over selected periods of time. They selected four different time-span according to the historical events at the time such as years after civil war, economic depression, Great Depression, etc. During the selected time-periods economy and cotton were the major matter of concern which is why all mention of the topic “cotton” were also saved onto the data corpora.

In order to avoid potential errors during the topic modelling, pre-processing steps such as spelling correction were done. To perform the topic modelling, UMASS Amhert's Machine Learning for LanguagE Toolkit (MALLET) was used.
The goal of the study was to have a good analyze of the output of the topic modelling, which is why the analyzes process was done by a historian. According to this historian, the result topics found by MALLET were useful and interesting. The quality and the selling of cotton crops were a highly recurrent topic with words such as "good", "middling", "ordinary", "crop", "bale" or "market". This result was expectable contrarily to the topic containing "houston April general hero san" at the time-span from 1865 to 1901 as according to historians the Battle of San Jacinto which marked the independence of Texas from Mexico was not remembered until the beginning of the twentieth century. The result of this topic modelling, however, proved they wrong as the battle was widely documented in newspapers even thirty years after the battle. (Yang et al., n.d.)

This LDA application example emphasizes the importance of the analyze of the output of the topic modelling and its contextualization. Therefore, knowing the subject of the news on which the topic modelling is applied is mandatory in order to make the most accurate analyzes of the results. When analyzing the results of the LDA applied to climate change new stories, it is important to know the notable things that may affect news coverage and to be able to associate the topics with facts and studies. The part 2.2.2. gives an idea of what could be expected and what could be used in the analyzes of the LDA's output.

2.2.2 Public view on climate change

Many organizations have made surveys over the years to find how big the concern of the public was towards climate change which were used afterwards in some studies. In 2012, Brullem et al published a study about the public concern over climate change in the US and the same kind of study was done in 2006 by Lorenzi and Pidgeon considering the European perspective as well.

In 2004 based on the surveys, 62% in the United Kingdom described climate change as a "fairly bad thing" or a "very bad thing" however 10% of the respondents considered it as a "fairly good thing" or a "very good thing" (Lorenzoni and Pidgeon, 2006). In the same year in the United States only 26% respondents of a survey worried “a great deal” about global warming (Brulle et al., 2012). In 2004 as well, an European survey of public opinions showed that environmental issues was the third biggest concern in the EU-15 with 45% of the respondents responding that (Lorenzoni and Pidgeon, 2006).

These percentages can be explained by the lack of knowledge in the subject by the population. A survey done in 1999 showed that in 27 countries very few knew that burning of fossil fuel impacted global warming, the highest percentage of people knowing was in Finland with 17% and the lowest in the US with only 11%. (Lorenzoni and Pidgeon, 2006)

The best ways to increase population’s knowledge on climate change is by giving better access to the information related to it. According to the study about shifting public opinion on climate change (Brulle et al., 2012) public’s concern is influenced by politics, major weather changes but mainly by media coverage even though media coverage doesn’t affect public opinion for more than a month and a half.

According to the previous information, we could expect that finding new stories on climate change around 2004 in the US may be harder than finding some in Europe in the same year. However, there is probably less new stories worldwide before 2000 as the lack of knowledge related to climate change is higher according to the percentages. We could also expect that the new stories will give more information about climate change when there are major changes in weather for example higher or lower temperatures or natural disasters such as wildfires or floods.

3 RESULTS

3.1 The program

The corpus of this study is composed of roughly 100 news articles from both New York times and The Guardian published since 1989. In order to achieve more precisely the goal of this study and to find the way news articles have presented climate change throughout the years, we decided to divide the corpus. The overall corpus is divided in the way that new articles from a timespan of 5 years are put together in csv files.

The LDA program in this paper is made using Susan Li's "Topic modeling and Latent Dirichlet Allocation" as a major help, as her code is modified in the way that it is usable with a smaller corpus and to achieve the goal of this study.

In order to perform properly an LDA-program, it is important to pre-process the dataset which involves tokenization, lemmatization and stemming. By tokenizing the dataset, we remove all the punctuation of the dataset and lowercase every word. The lemmatizing process is when the program puts every word into first person and verbs into present. According to Jacobi et. al. in their LDA study on nuclear energy, the lemmatizing process completes the stemming process as in the stemming process some words may have different stems even if they are conjugated from the same verb (Jacobi et al., 2016). Finally, in the stemming process the program reduces every word into their root by suppressing prefixes and affixes from words. In this program the stemmer used is the improved Porter stemmer known as Snowball Stemmer.

According to Wiese et al. Porter stemmer is one of the most used stemming algorithms as it is simple and quick to use, but as a result some words are not stemmed correctly (Wiese et al., 2011). Porter says that the porter algorithm has faced three major problems since its creation such as the will to improve it, bugs in the code and misunderstanding of the initial algorithm. To overcome these problems, he created Snowball stemmer which is more accurate than its predecessor.

The stemming in the program works thanks to the use of the following code:

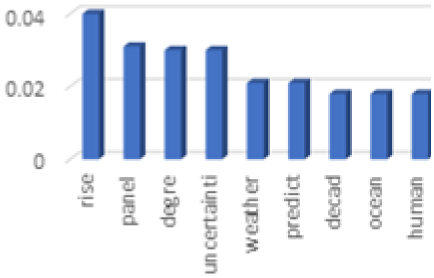
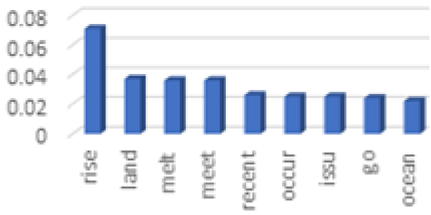
```
stemmer = SnowballStemmer('english')
def lemmatize_stemming(text):
    return stemmer.stem(WordNetLemmatizer().lemmatize(text,pos='v'))

def preprocess(text):
    result = []
    for token in gensim.utils.simple_preprocess(text):
        if token not in gensim.parsing.preprocessing.STOPWORDS and 
len(token)>3:
            result.append(lemmatize_stemming(token))
    return result
processed_docs = documents['text'].map(preprocess)
```

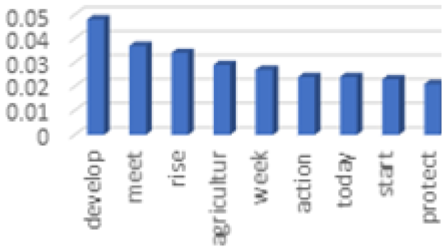
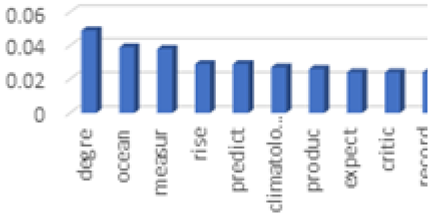
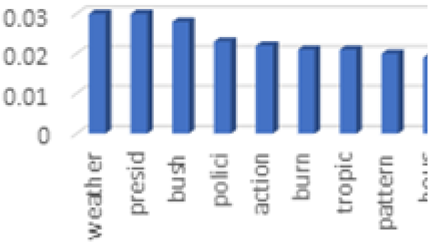
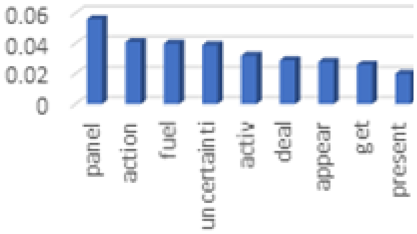
In order to filter out some tokens so that the result won't have words that are not relevant for the corpus, we decided to not include tokens that are found in less than three documents of the corpus. The number is chosen based on the size of the corpus which is small compared to the average LDA studies. In this study we have about a hundred news articles which are divided into six different corpuses and every smaller corpus is only containing about ten news articles. Moreover, we decided to keep out of the result the tokens occurring in more than 50% of the entire corpus as they might be words that are from daily vocabulary. As the six corpuses are small, it was decided to keep only 100 most relevant tokens from each corpus.

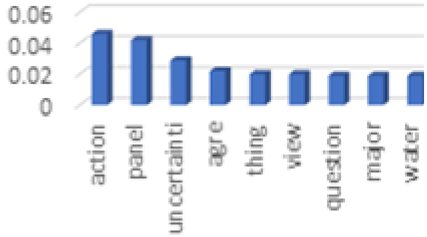
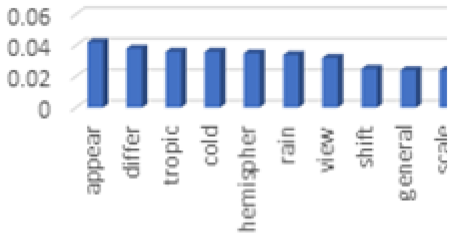
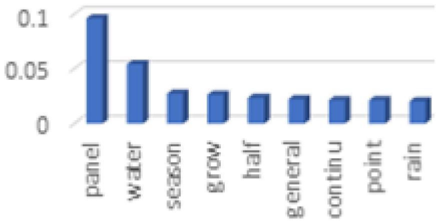
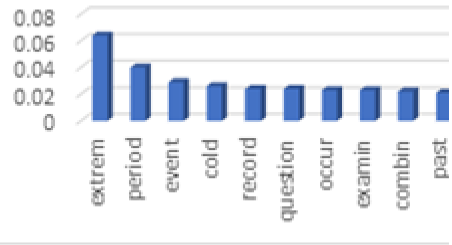
The program gives as an output a list of five topics with ten words along with their frequency. As previously mentioned, the corpuses used in this study are small which is why the number of topics is so small.

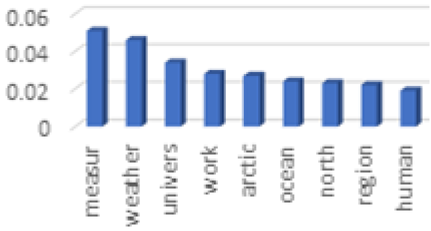
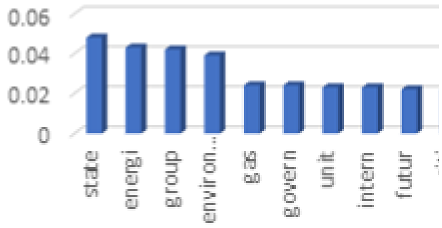
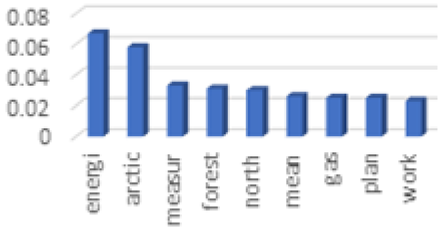
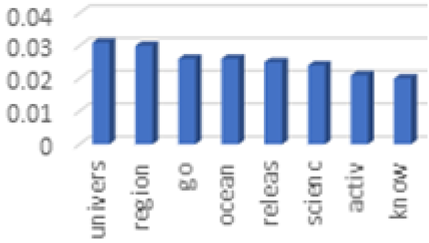
3.2 The program output

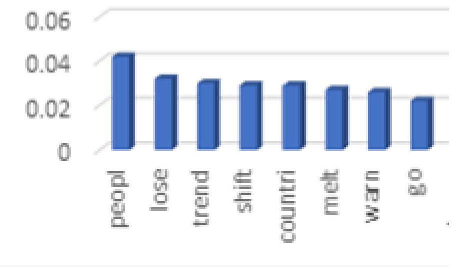
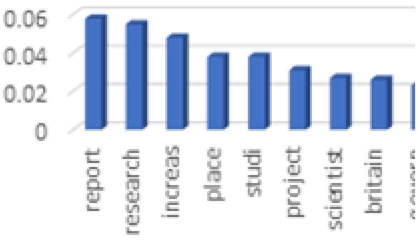
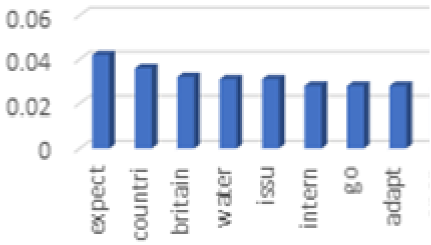
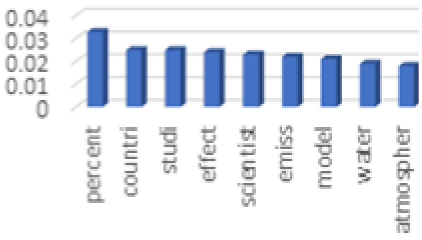
Table 1 Most occurring words in the corpuses by topic

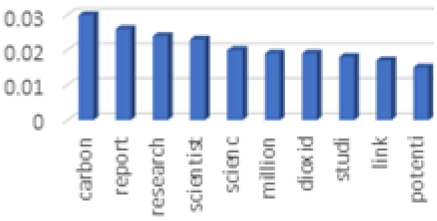
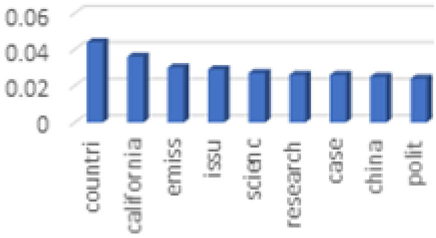
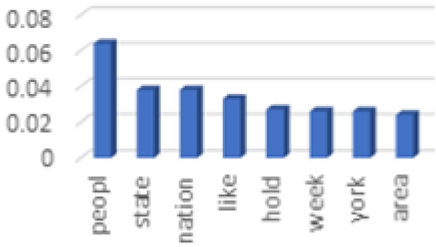
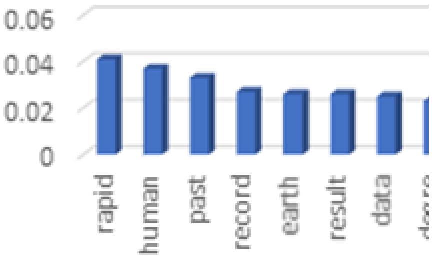
TIME-SPAN		**MOST OCCURRING WORDS**	**GRAPHICAL REPRESENTATION**
1989 -1994			
	Topic 1	Rise Panel Degre Uncertainti Weather Predict Decad Ocean Human Step	
	Topic 2	Rise Land Melt Meet Recent Occur Issu Go Ocean Tropic	
	Topic 3	Develop Meet	

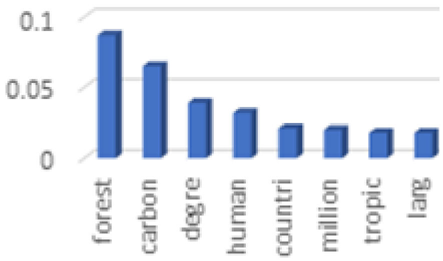
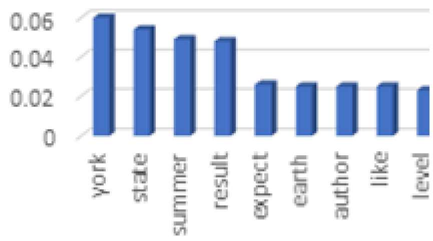
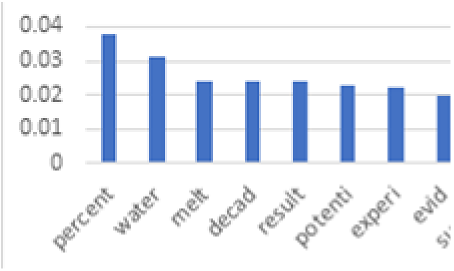
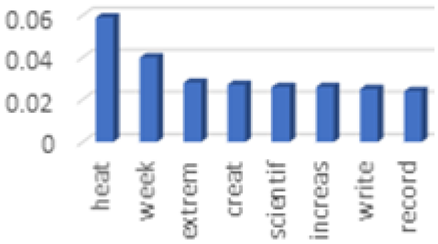
		Rise Agricultur Week Action Today Start Protect Intern	
	Topic 4	Degre Ocean Measur Rise Predict Climatologi Produc Expect Critic Record	
	Topic 5	Weather Presid Bush Polici Action Burn Tropic Pattern Hous Long	
1995 -1999	Topic 1	Panel Action Fuel Uncertainti Activ Deal Appear Get Present Today	
	Topic 2	Action Panel	

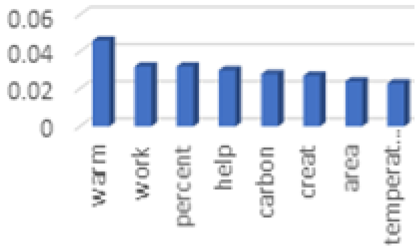
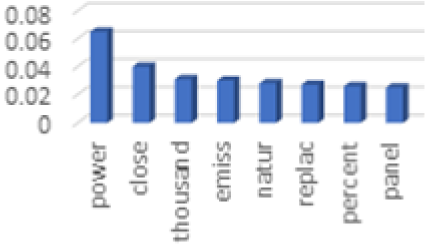
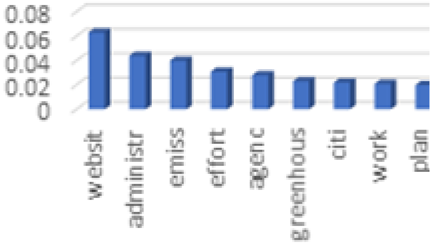
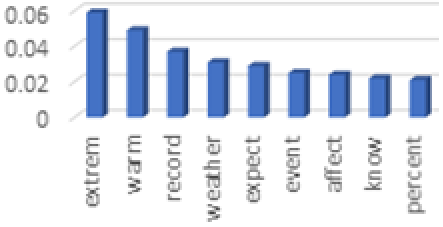
		Uncertainti Agre Thing View Question Major Water Flood	
	Topic 3	Appear Differ Tropic Cold Hemispher Rain View Shift General Scale	
	Topic 4	Panel Water Season Grow Half General Continu Point Rain Climatologi	
	Topic 5	Extrem Period Event Cold Record Question Occur Examin Combin	
2000 - 2004	Topic 1	Measur Weather Univers Work	

	Arctic Ocean North Region Human Peopl	
Topic 2	State Energi Group Environment Gas Govern Unit Intern Futur Citi	
Topic 3	Energi Arctic Measur Forest North Mean Gas Plan Work Help	
Topic 4	Univers Region Go Ocean Releas Scienc Activ Know Degre Environment	

2005 - 2009	Topic 5	Peopl Lose Trend Shift Countri Melt Warn Go Know Bring	
	Topic 1	Report Research Increas Place Studi Project Scientist Britain Govern Futur	
	Topic 2	Expect Countri Britain Water Issu Intern Go Adapt Open Flood	
	Topic 3	Percent Countri Studi Effect Scientist Emiss Model Water Atmosphere Region	
	Topic 4	Carbon	

2010 - 2014		Report Research Scientist Scienc Million Studi Link Potenti	
	Topic 5	Countri California Emiss Issu Scienc Research Case China Polit Requir	
	Topic 1	Peopl State Nation Like Hold Week York Area West Focus	
	Topic 2	Rapid Human Past Record Earth Result Data Degre Percent Higher	
	Topic 3	Forest Carbon Degre	

<table>
<tr><td></td><td></td><td>Human
Countri
Million
Tropic
Larg
Part
Decad</td><td>
</td></tr>
<tr><td></td><td>Topic 4</td><td>York
State
Summer
Result
Expect
Earth
Author
Like
Level</td><td>
</td></tr>
<tr><td></td><td>Topic 5</td><td>Percent
Water
Melt
Decad
Result
Potenti
Experi
Evid
Suggest
Question</td><td>
</td></tr>
<tr><td><b>2015 - 2019</b></td><td>Topic 1</td><td>Heat
Week
Extrem
Creat
Scientif
Increas
Write
Record
Weather
Make</td><td>
</td></tr>
<tr><td></td><td>Topic 2</td><td>Warm
Work
Percent
Help
Carbon</td><td></td></tr>
</table>

	Creat Area Temperatur Environment Know	
Topic 3	Power Close Thousand Emiss Natur Replac Percent Panel Organ Plan	
Topic 4	Websit Administr Emiss Effort Agenc Greenhouse Citi Work Plan Environment	
Topic 5	Extrem Warm Record Weather Expect Event Affect Know Percent Temperatur	

<u>1989 – 1994</u>:

During the first years analyzed in this paper, the main ideas coming through new articles according to LDA are the newness of climate change with word like “occur”, “recent”, concern about the rise of sea level and lack of knowledge around ocean and weather-related changes. Words like “critic” and “record” are clearly showing that the changes in climate are getting bigger and that the concern is real. The word “panel” is also a frequent word in this time-span and it may refer to IPCC which was created in 1988. Finally, it is important to note the presence of words like “predict, “climatologi” and “expert” which shows that information relayed by newspaper are backed up.

1995 – 1999:

The word “panel” is occurring in three topics out of five and we will suppose as for the topic 1 of years 1989 to 1994 that it is referring to IPCC. Moreover, the doubts are still present as there are many words referring to questioning such as “uncertainti” and “question”. The main concern in this time-span is about water probably because some major floods occurred around the world in 1999. Some extremes and records are also beaten during this period such as heavy rains or cold weather as it can be seen by words like “cold”, “record” and “extrem”.

2000 – 2004:

In this time-span words referring to humans are much more recurrent than in the previous two life-span which shows that the news-articles are giving the idea that people are influencing the climate change in some way. The word “gas” is also mentioned as well as “city” which shows that the new articles are suggesting that the climate change may be affected by these human related things. The arctic and north region in general is also an emerging topic in these years which means that the consciousness about the effects of global warming on the arctic sea and north region are being awaken.

2005 – 2009:

In these years the news articles were for some part about water but the most important thing from that period is the scientific research. As we can see in table 1, many words from the topics are about science, Scientifics, studies, projects and research. This means that in this period, newspapers gave the idea to people of a serious problem as scientific are working on it and there is some concrete data to be released to prove that climate change is real.

2010 – 2014:

During these years newspapers used words such as “data”, “percent”, “result” and “expect” which proves there is less incertitude than in 1989 – 1999. There are also many words referring to greatness like “million” and “large” which shows that the concern is getting bigger. These words could also refer to an increase in the impacts of global warming.

2015 – 2019:

In the recent years, newspaper have used the word “environment” much more as it has become a major issue. Incertitude is not present at all. Words that we know today to create greenhouse effect are mentioned like “carbon” and “emission”. In these years the words showing some extremes are still present such as “extrem” and “record”. The word “increase” alongside “temperature” shows that temperature is getting warmer. We can also see the word “warm” coming up which is referring to “global warming” and shows that global warming is in the newspaper as a fact and not something we are not really sure about.

Based on the previously stated analyzes timespan by timespan we can see that the media coverage over climate change has changed in some ways for the past 30 years. Even if the concern has surprisingly been present since 1989, the uncertainty about what climate change really is and its consequences was much more present at the beginning than at the last time-spans. Before 2000 the main concern was the changes in ocean and in weather but after that the newspapers started to relay information about what is causing climate change and how big the issue is becoming.

4 DISCUSSION

4.1 Result analyses

The best way of analyzing the result is going on the same line than Jacobi et. al. on their study on the evolution of opinion over nuclear energy in new stories or than Yang et. al. on their study on historical newspapers, which is making a parallel with events and studies around the same timespan as the corpus.

In the beginning of this paper, we supposed that major natural disasters such as wildfires or floods could affect the way newspaper are writing about climate change but surprisingly words like "flood" were not encountered around 2004 even though there were some important natural disasters in Asia at this time. Finding news articles in the US was not harder than finding some in Europe about climate change around 2004.

In the result of this program, skepticism from 1989 hasn't been found at all but only uncertainty about the climate change has. Even though words referring to skepticism haven't been found, the uncertainty around climate change relayed by newspapers could affect people's opinion in a skepticism increasing way. As Riley E. Dunlap mentions in his study about climate change skepticism, an important part of the population can be skeptic towards global warming because its causes, issues and solutions are unclear or not known (Dunlap, 2013). As between 2005 and 2009 the topics are composed of more words referring to Scientifics and concrete data, the information given by newspaper from that time on are reducing public skepticism.

The International Panel on Climate Change was created in 1988 (Bolin, 2008), which is one year before the beginning of this study. This explains why the concern about climate change was present in news stories since 1989. Furthermore, Bert Bolin says that the implication of human in global warming became a major concern during the twentieth century (Bolin, 2008) which explains why there were news stories on climate change even in 1989.

According to IPCC, eleven of the twelve years between 1995 and 2006 were among the twelve warmest years known since 1850 to 2007 based on the global surface temperature. This increase in the temperature is partly caused by anthropogenic greenhouse gases such as carbon dioxide which concentration in the atmosphere exceeded its natural range in 2005 with a range of 379 ppm^3 instead of a natural range of 280 ppm^3. This concentration has increased the most since 1960 between 1995 and 2005 with an average growth of 1.4 ppm per year. (Solomon et al., 2007) This explains the fourth topic between 2005 and 2009 with the word "carbon" occurring at the higher frequency alongside words referring to science.

From 1989 to 1994 we can see that many of the topics contain the word "rise" and that water related word are frequent. Other important words at this time span are related to temperature and melting. These topics are explained by the physical science basis on climate change of 2007, which shows that the sea level rose from 1993 to 2003 at a rate of 3.1 mm per year partly due to the melting of ice sheets of Greenland and Antarctica as well as melting in glaciers and ice caps (Solomon et al., 2007).

According to the results, the financial issues related to climate change were not relayed by newspaper and neither did the changes in biodiversity, which are some important topics today. Some important topics were not found as well such as the new laws regarding single-use plastic or decisions made by politicians regarding global warming.

Even though some important topics were not visible in the result of this LDA-program, we could still make suppositions on how news stories will change in the future. In this study, we found that the incertitude decreased throughout the years which could make us except that newspapers will relay more study-related and science-related articles in the future to prove that climate change is real. The concern about climate change has also only increased in 30 years which proves that it will continue to do so and maybe finding solution to reduce or to eliminate global warming will be the main topic in a few years.

4.2 The effect of the program on the result

In the study lead by Jacobi et. al. about news stories on nuclear energy, they presented two ways of defining the numbers of topics in the LDA model: mathematically or by human-interpretation. In mathematical one, the perplexity is defined by running the LDA program on a small portion of the corpus and then by evaluating the model with the result. The human-based consists of evaluating the interpretability based on the corpus. (Jacobi et al., 2016) In this study we decided to use the latter one as the corpus is small and five topics for a corpus of roughly ten documents is good as we can see in the results that some topics are already containing many same words as others on the same corpus. If the number of topics was increased it could give non-relevant and repetitive topics as an output because the filtering was made in the way that no words appearing in less than 3 documents or more than 50% of the corpus are included in the finale result.

By increasing the number of documents where a same word is appearing the topics received as output wouldn’t be varied enough as there are only ten documents in each corpus. However, if the corporas were bigger, it would be mandatory to adapt this number to the corpus size as well as adapting the number of topics. Even if the number of documents where a word needs to appear is small in the program, it could still affect the result due to the small size of the corpus. In fact, some important topics from the recent year were missed, even though they were widely presented in news stories. Such topics are for example the young climate strikers who are skipping schools since 2016 to protest for the environment based on the action of Greta Thunberg or the retreat from the Paris agreement of the United States in 2017. The absence of these topics from the result of the program could also be explained by the small size of the corpuses and news stories on these two topics were not necessarily included in the corpus used in the program.

Filtering out words occurring in more than 50% of the entire corpus may affect the result and let out some important words. This may explain the absence of words associated with natural disasters as they might have been widely documented during precise timespans and the result might be more precise if the percentage was decreased to 30% for example. Decreasing the percentage could still leave too much unwanted words in the results as it may leave common words such as verbs like “have” or “be”.

As Jacobi et. al. noted in their study, lemmatizing and stemming is important even though lemmatizing is not necessarily mandatory with English as it is a simple language. In this study it was decided to still perform it as the result is more precise than if we decided to use only stemming.

The stemmer used in the program of this study as previously stated is the snowball stemmer which is considered as a light stemmer and not the best of them. Both Porter stemmer are indeed outperformed by KStem developed by Krovetz. (Wiese et al., 2011) However, the snowball stemmer is easier to use thanks to its lightweight and to its presence in the nltk library of Python which makes it easy. It also fulfills perfectly the role of a stemmer in this study as the corpuses are not wide. For a wider corpus, the use of another stemmer would be more efficient, more precisely a dictionary-based stemmer. An example of a dictionary-based stemmer is the Hunspell stemmer which is used as a spell checker in Open Office, LibreOffice, Chrome and Firefox but is case sensitive, due to that it lowercasing the characters of the corpus beforehand will give a better result (Hunspell Stemmer | Elasticsearch).

5 REFLECTIONS AND FUTURE WORK

The study lead in this article is not as wide as a major part of LDA studies where corpuses are composed of many thousands of news articles which have affected the result. For a more accurate result the corpus should be wider to allow the LDA program to find more topics throughout the years and to have a better idea of how things have changed. Moreover, the main source of news articles for the corpus used in this study was New York Times but finding more newspapers that have made articles about climate change since 1989 throughout the world could give a wider answer to this study's problematic. Natural disasters didn't, according to the corpuses analyzed in this paper, affect news stories but we could suppose that not enough news articles were about natural disasters as there were only 5 by year. In order to achieve a more precise result and to create wider corpuses coding a program that will fetch news stories by itself from the web would be great but due to lack of time and of current knowledge this was not possible yet. Increasing the corpus size could also be easier by using newspapers' APIs with the word "climate change" between 1989 and 2019 as Jacobi et.al. did in their study with the New York times API.

Developing this study by changing the timespan would be interesting as well. Making the study start at the middle of the 20th century for example could give a better idea of the evolution of skepticism. The consciousness in this time was not as important as it is today and the development of industry and technologies at that time were more important than the impacts on the environment. As according to IPCC, the earth's surface temperature started to increase in 1965 at a higher speed than before (Solomon et al., 2007). However, finding enough news stories to perform a proper topic modelling algorithm on them could be hard as the digitalization started only in the end of the 20th century and not all the newspaper are digitalized. Another problem that could be faced is the lack of measures and data from the middle of 20th century which may have reduced the amount of news stories due to lack of information to relay at that time.

Even though in this study we decided to use English as the only source language for the news stories, creating a multilingual program which could find topics from documents written in different languages would give a much more precise output. It would indeed give a precise idea of how the viewpoint on climate change has changed not only in

English-speaking countries but worldwide and consequently finding skepticism would be easier.

Achieving the goal of using text sources in numerous languages could be achieved by using a JointLDA, where a layer of hidden variables is added to the basic LDA algorithm which are called concepts. The goal of JointLDA is indeed to find topics in cross-lingual corporas. The concepts in JointLDA are dictionary entries and in this case the dictionary is bilingual. However, as the bilingual dictionary is never as wide as a monolingual one, some additional work needs to be done on the dictionary by adding artificial entries. (Jagarlamudi and Daumé, 2010)

6 CONCLUSION

In this study we used LDA to find how viewpoint in news stories has changed from 1989 until today. Based on the results, the main idea coming through is the evolution of the incertitude towards climate change during the last 30 years. At the end of 20th century the incertitude was much more important than today due to the lack of knowledge around global warming, but from the beginning of the 21st century scientific data and research became more present. Moreover, the creation of IPCC has influenced the news stories since its creation in 1988. Some major climate change related events came through the topic modelling such as the rise of sea level, the rise of temperature and research on climate change. In the future we could except news stories to increase the awareness of their reader's on climate change by relaying scientific studies or reports. We could also except that news stories will keep giving information about global warming's impacts on the environment. It could be interesting to perform the same study with different parameters and bigger corpuses in order to find more precise results and clearer ideas on how the viewpoint has evolved since 1989.

REFERENCES

Bíró, I., Szabó, J., Benczúr, A.A., 2008. Latent Dirichlet Allocation in Web Spam Filtering, in: Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web, AIRWeb ’08. ACM, New York, NY, USA, pp. 29–32. https://doi.org/10.1145/1451983.1451991

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022.

Bolelli, L., Ertekin, Ş., Giles, C.L., 2009. Topic and Trend Detection in Text Collections Using Latent Dirichlet Allocation, in: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (Eds.), Advances in Information Retrieval, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 776–780.

Bolin, B., 2008. A History of the Science and Politics of Climate Change by Bert Bolin. Cambridge university press. https://doi.org/10.1017/CBO9780511721731

Brulle, R.J., Carmichael, J., Jenkins, J.C., 2012. Shifting public opinion on climate change: an empirical assessment of factors influencing concern over climate change in the U.S., 2002–2010. Climatic Change 114, 169–188. https://doi.org/10.1007/s10584-012-0403-y

Carlsson-Kanyama, A., 1998. Climate change and dietary choices — how can emissions of greenhouse gases from food consumption be reduced? Food Policy 23, 277–293. https://doi.org/10.1016/S0306-9192(98)00037-2

Dow, K., Downing, Thomas E, 2006. The atlas of climate change : mapping the world’s greatest challenge. Earthscan.

Dunlap, R., 2013. Climate Change Skepticism and Denial: An Introduction. American Behavioral Scientist 57, 691–698. https://doi.org/10.1177/0002764213477097

Fei-Fei, L., Perona, P., 2005. A Bayesian hierarchical model for learning natural scene categories, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05). Presented at the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), pp. 524–531 vol. 2. https://doi.org/10.1109/CVPR.2005.16

Hong, L., Dan, O., Davison, B.D., 2011. Predicting Popular Messages in Twitter, in: Proceedings of the 20th International Conference Companion on World Wide Web, WWW ’11. ACM, New York, NY, USA, pp. 57–58. https://doi.org/10.1145/1963192.1963222

Hunspell Stemmer [WWW Document], n.d. URL https://www.elastic.co/guide/en/elasticsearch/guide/current/hunspell.html (accessed 6.5.19).

IPCC presents findings of the Special Report on Global Warming of 1.5°C at event to discuss Viet Nam’s response to climate change — IPCC, 2018. URL https://www.ipcc.ch/2018/10/10/ipcc-presents-findings-of-the-special-report-on-global-warming-of-1-5c-at-event-to-discuss-viet-nams-response-to-climate-change/ (accessed 3.10.19).

Jacobi, C., Atteveldt, W.H. van, Welbers, K., 2016. Quantitative analysis of large amounts of journalistic texts using topic modelling. Digital Journalism 4, 89–106. https://doi.org/10.1080/21670811.2015.1093271

Jagarlamudi, J., Daumé, H., 2010. Extracting Multilingual Topics from Unaligned Comparable Corpora, in: Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., van Rijsbergen, K. (Eds.), Advances in Information

Retrieval. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 444–456. https://doi.org/10.1007/978-3-642-12275-0_39

Lorenzoni, I., Pidgeon, N.F., 2006. Public Views on Climate Change: European and USA Perspectives. Climatic Change 77, 73–95. https://doi.org/10.1007/s10584-006-9072-z

Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T., 2005. Discovering object categories in image collections, in: ICCV 2005.

Smith, I.J., Rodger, C.J., 2009. Carbon emission offsets for aviation-generated emissions due to international travel to and from New Zealand. Energy Policy, New Zealand Energy Strategy 37, 3438–3447. https://doi.org/10.1016/j.enpol.2008.10.046

Solomon, S., Intergovernmental Panel on Climate Change, Intergovernmental Panel on Climate Change (Eds.), 2007. Climate change 2007: the physical science basis: contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge ; New York.

Stanley, J.K., Hensher, D.A., Loader, C., 2011. Road transport and climate change: Stepping off the greenhouse gas. Transportation Research Part A: Policy and Practice, A Collection of Papers:Transportation in a World of Climate Change 45, 1020–1030. https://doi.org/10.1016/j.tra.2009.04.005

Wiese, A., Ho, V., Hill, E., 2011. A comparison of stemmers on source code identifiers for software search. pp. 496–499. https://doi.org/10.1109/ICSM.2011.6080817

Xing, D.S., Girolami, M., 2007. Employing Latent Dirichlet Allocation for fraud detection in telecommunications. Pattern Recognition Letters 28, 1727–1734. https://doi.org/Xing, D. and Girolami, M. <http://eprints.gla.ac.uk/view/author/9221.html> (2007) Employing Latent Dirichlet Allocation for fraud detection in telecommunications. Pattern Recognition Letters <http://eprints.gla.ac.uk/view/journal_volume/Pattern_Recognition_Letters.html>, 28, pp. 1727-1734. (doi:10.1016/j.patrec.2007.04.015 <http://dx.doi.org/10.1016/j.patrec.2007.04.015>)

Yang, T.-I., Torget, A., Mihalcea, R., n.d. Topic Modeling on Historical Newspapers 9.

Yano, T., Cohen, W.W., Smith, N.A., 2009. Predicting Response to Political Blog Posts with Topic Models, in: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL '09. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 477–485.