University of Bergen

Measuring opinions about climate change:

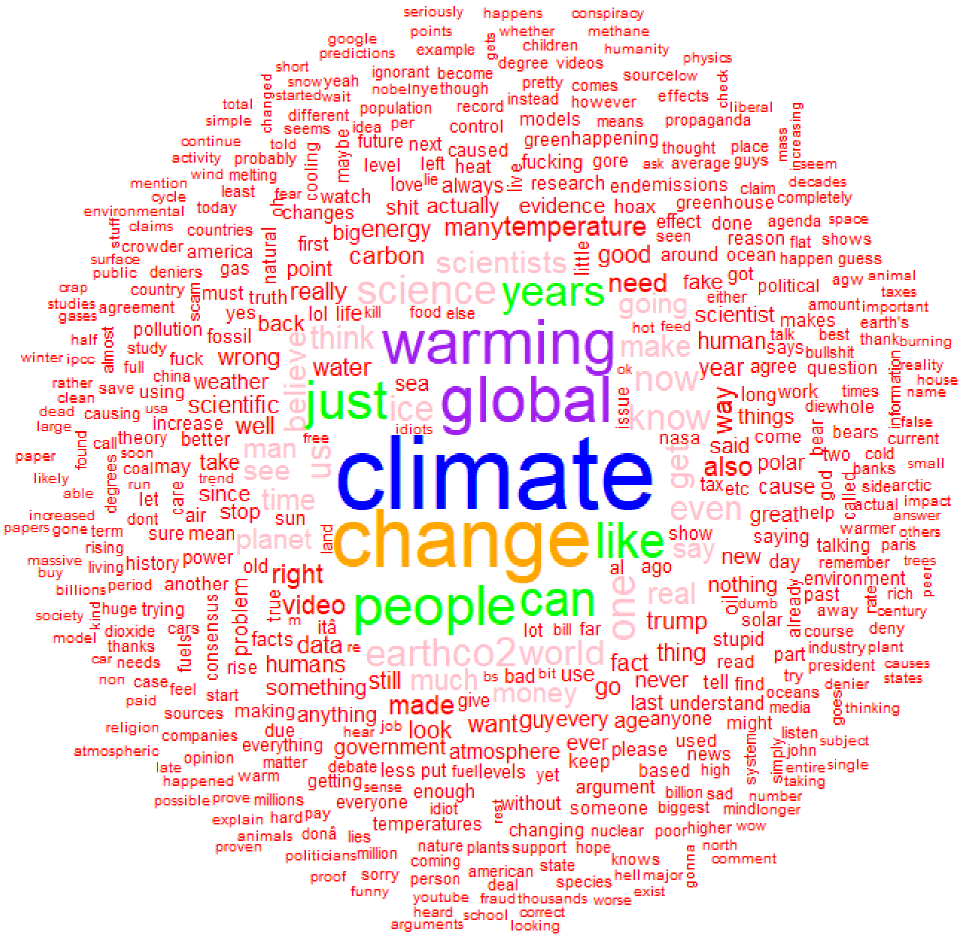
The Method of Nonparametric Automated Content Analysis applied

A thesis submitted in fulfillment of the requirements
for the degree of Master of Informatic Science

in the

Department of Information Science and Media Studies
Faculty of Social Sciences

Author:
José de Jesús Martinez Guardiola
Bergen, Juni 2019


climate
change
global
warming
people
just
like
years
can
one
earthco2world
science
scientists
know
now
make
think
believe
time
planet
money
much
temperature
energy
carbon
good
need
video
right
made
trump
real
even
also

Abstract

Today the importance of automated methods of content analysis is gaining more relevance thanks to the exponential increase of digital information in recent years. Currently, we can find digital content on the Internet that is relevant for researchers in diverse areas. The opinions that social media users post on the Internet are of substantial importance for politicians and scientists in social science, since they can use this information to study the behavior of these groups of people or, simply in order to learn about the general opinion about a specific topic.

This thesis introduces the automated content analysis as a field of science and shows how to implement the method of Hopkins and King. This method calculates the proportions of categories in texts in order to identify feelings, grade of opinion or simply to classify among multiple topics. Hopkins and King's method gives an advantage of being able to directly calculate these category proportions without depending on other filtering methods. In previous works it has been verified that this method presents a better performance regarding the calculation of the proportions of categories and in comparing them with other direct classification methods or parametric methods for text analysis.

Around 83,000 opinions from YouTube videos will be analyzed with this method. This study measures the performance of the nonparametric method implemented in opinions related with the climate change. The method will be used to measure the percentage of the total number of comments made by climate change activists and deniers.

Acknowledgements

This thesis would not have been possible without the support of my friends and family. would like to thank my friend Ragnhild Nyheim. I am forever gratefully indebted to her for her positive support and language assistance.

I would also like to thank my family: My mother Maria Inocencia and my daughter Aile Maria. Thank you for your love and inspiration. I am also very grateful to Aile Marias mother, Ellen, for her cooperation and support.

Finally, I want to express my very profound gratitude to my future wife Kristine. Thank you for your encouragement, love and understanding.

José de Jesús Martinez Guardiola

Contents

List of figures

List of Tables

Abbreviations

HK We are using this abbreviation for refer to the method of Hopkins and King's nonparametric method for automated content analysis

SA Sentiment Analysis

ACA Automated Content Analysis

ATA Automated Text analysis

ML Machine learning

SVM Support Vector Machines

OM Opinion mining

1. Introduction

Currently, almost all information is produced digitally every day around the world. The excessive growth of information does not stop, and the information stored in digital devices will be increasing year after year (Barbosa & Aoki, 2009). Blogs and social media platforms are examples of the constant growth of the data generated. On Facebook and Twitter millions of people share their opinions about different topics or they share their daily activities with friends or family (Faizi, El Afia and Chiheb, 2013). All this data is important information that can be used by researches, politicians, journalists, social scientists or enterprises to learn about the opinions of people on a specific topic (Pang and Lee, 2008).

The need to estimate proportions of opinions on a specific topic is a subject of great interest to politicians and scientists. Due to the large number of data available on digital sources, it is essential to use methods of automatic content analysis. These kinds of processes are not perfect and may present biases, and thus there is an increasing need for better methods in order to classify and organize large quantities of data. The field of automated content analysis brings indispensable approaches for the management of these large data collections. Digital content comes from text, audio, videos files, or any format of images stored in electronic media such as e-books, blogs or social media. In this thesis the main focus will be to analyze opinions from texts extracted from social media.

In this thesis I will measure the performance of the nonparametric method for automated content analysis (Hopkins & King, 2010) in order to estimate the proportion of documents classified into a specific category on the topic of climate change.

Regarding climate change, the two main tendencies are those who present opinions that agree with the fact that the current climate is caused mainly by human activity and those who think it is not correlated. The opinion of people in social media about the climate change will be tracked to determine if the quantity of opinions agrees or not with the question: Is the climate change mainly caused by human activities?

Even though scientific evidence shows that the current climate is changing due to human activities, there is also a belief that denies that climate change is caused by humans and these individuals refuse to take actions to reduce the greenhouse gas emissions. With this research I intend to measure the performance of the Hopkins and King method when analyzing opinions about climate change extracted from YouTube.

1.1 The problem statement

The existing methods to estimate proportions are direct sampling and aggregation of individual document. They could present biases calculating category proportions when the sample is non-random, contemplating that in most classification problems the data is not random (D. Hopkins and King, 2010a). The nonparametric method has been proved as a good alternative to calculate proportions with nonrandom data presented. (King and Lu, 2008; D. Hopkins and King, 2010b; Ceron, Curini and Iacus, 2014)

The Hopkins and King method has not yet been tested for the topic of climate change working with opinions from social media. Some times, these kinds of opinions are difficult to understand for humans because they contain different levels of language; from well-written text to informal language or even slang, and there are also sarcasms or opinions that are badly formulated. “Sentiment categorization is difficult to realize because of the mixed data types and because the language used vary from well written language to colloquial not-well used languages”(D. Hopkins and King, 2010, p231). According to Pang, Lee and Vaithyanathan, sentiment categorization is more difficult than topic classification (Pang, Lee and Vaithyanathan, 2002) .

1.2 The climate change as a problem statement

Scientists and institutions continually publish videos on social media about the effects of the climate change in order to create a positive impact on the viewer in order to reduce the human activities that cause climate change. The opinions about the videos make a direct impact on the viewers. For scientists and for publishers in general, it is of great interest to know the proportions of positive and negative comments about their publications.

Videos related with climate change are polarized. On one side, there are have the videos that are published by institutions, scientists, or activists. While the other side, there are videos that are published by deniers that pretend to demonstrate that the climate change is just a hoax, a campaign to collect taxes or simply a conspiracy. To measure the polarity in opinions I first need to classify the videos before I classify the opinions. As part of the methodology I will categorize the videos in favor or in disfavor of human made climate change, before I categorize the opinions. It is important to avoid classifying opinions like “this is a farce”, “it is not true” or “please present evidence” because these kind of opinions are present in both kind of videos.

1.3 Motivation

When the goal is to measure the general opinion of climate change, one should consider the opinions on the Internet as a source of data, this to avoid time consuming and expensive surveys. One can extract large quantities of data from the Internet that can present the opinions needed for the study. Machine learning is a tool to automatize the analysis of huge volume of opinions in form of unstructured text.

The increasing number of available data in digital media represents an opportunity to use this information for different kinds of researches. In these kinds of studies, human effort would be insufficient to process such large volumes of data. The method of Hopkins and King is proved to be an excellent tool to estimate proportions from large sets of documents with the minimal human effort. Just a small subset of documents manually labeled is necessary to determine the category proportions. (King, 2007, 2016; D. J. Hopkins and King, 2010).

Exploring new variants to implement this method could provide useful information about how to use this method, its performance and the challenges working with YouTube Comments. This is an exploratory study where the objective is to find the best performance of the method doing experiments with diverse data and different procedures for text mining.

1.4 Research Questions

Does the nonparametric method perform well with the topic of climate change?

I will test the method using opinions related with climate change, then I will measure the results comparing with the human labeled sets.

The opinions from the YouTube platform is a reliable source of for text mining?

In a previous analysis of the opinions on YouTube, I found that some opinions extracted from YouTube was difficult to classify as pro climate change or against, in this study I will found the difficulties working with this kind of information for text mining and find the best way to tackle these difficulties.

2. Literature Review

This chapter explains the essential concepts needed in order to understand the nonparametric method and how to put it into practice. The concepts and theory about Content Analysis are introduced in this section as well as the Machine Learning approach. Other relevant algorithms are also explained in this literature review. It is, additionally, essential to know the different techniques and methods in this field used for the categorization of documents, sentiment analysis and text analysis. Concepts in probability and statistical theory are also reviewed. All these terms are frequently used in the literature of nonparametric method and in general in the Machine Learning literature.

2.1 Automated Content Analysis

Automated content analysis is a relatively new field in science that has been incorporated to facilitate the content analysis. It proves to be essential where large quantities of documents need to be processed. In this part of the literature review, I present an introduction to the content analysis as a field of science, as well as other studies in the field of automated content analysis.

2.1.1 Content Analysis

Content analysis studies the content with reference in the meaning, context or intentions in documents. It is used as a scientific tool to provide new insight, or to increase the understanding of a certain topic (Krippendorff, 2003; Prasad, 2008; D. J. Hopkins and King, 2010). The term Content analysis has been utilized for 75 years and has been described in Webster's Dictionary of English language since 1961 (Prasad, 2008).

Content refers to all kind of formats that brings relevant information. The content can be defined as texts, audio, videos or pictures. This thesis is mainly focused on the field of text analysis where the sources comes from unstructured texts in sources like books, essays,

interviews, discussions, newspaper headlines, articles, historical documents, speeches, written conversations, advertising, theater, informal conversations, or really any occurrence of communicative language in the form of a text. In this thesis, content analysis will be focused on the use of unstructured text extracted from opinions posted on the YouTube video platform, in videos related with the climate change. YouTube is currently the most popular video platform, where users can share and watch videos. Users can interact with each other giving a like or dislike to the posted videos (YouTube, 2019). They can also interact by giving written opinions related to the content of the video. For this research, these opinions will be the study content to test the method of Hopkins and King, in order to measure the public opinion about climate change.

Content analysis is a time-consuming process. Historically, all the analysis were done manually (UMSL, 2004). Today we can make use of this tool using computers and software that process a large amount of data. However, the human factor is still indispensable due to the complexity of language. For this reason, the nonparametric method is presented as a good alternative to calculate document category proportions. This is a well-functioning method where the human factor is indispensable but without representing high costs or effort when analyzing large quantity of data.

2.1.2 The digital sources

The type of documents that are to be analyzed should be closely related to opinions about the chosen topic. Opinions can be found in all types of digital media such as blogs, social networks, news, transcripts of political debates, etc. These documents are usually in unstructured text format, but the method can also work with other kinds of digital content, such as photographs, audio or video.

Blogs and micro blogs such as Twitter, provide us with a very good source of information, since these usually express opinions, are in chronological order, and are widely accepted anywhere in the world. It is estimated that Facebook, for example, has around 2.3 billion active users monthly, while Twitter has around 323 million active users monthly. Blogs continue to

represent an important source of opinions with around 440 million blogs around the world. (J. Clement, 2019)

Opinions are extracted from blogs using techniques and information extraction tools and archived in standard documents formats for a later computer processing. The total population of documents is divided into two sets. One set will contain mostly the quantity of the documents; computers will process this part, while the other part will represent a random sample of the population, and this sample will be a small number of documents, which will be classified manually by social scientists.

2.1.3 The social aggregate

The social aggregate is a representation of people that frequently states their opinions in in blogs or social networks on the Internet. Opinions are strongly related to the topic of study. Depending on the target, the population of study could be activists, the media, the public opinion, elite influencers, etc. The methodology of this thesis contemplates people that post sporadically as well. They participate in the public conversation about the topic in question. As an example, there are persons who normally writes or blogs about everyday life like cars, gardening, food recipes among other things, but suddenly they will post a few opinions about climate change because is a trendy topic on the Internet. They also represent opinions than can influence other people to change the meaning in a relevant topic.

2.1.4 Automated content analysis

With the introduction of computers, content analysis has been automated, complementing other fields of science such as Machine Learning and text analysis and playing an integral role in the development of artificial intelligence. One advantage is that automated content analysis (ACA) let us reduce costs and time in data analysis producing qualitative and quantitative results. (Nunez-Mir *et al.*, 2016).

Context identification, concept definition and text classification are the most common tasks in the process of content analysis. These tasks are present in the methods to be analyzed in this thesis. In the later chapters we can see in detail its implementation. Another important task is the validation of the results. To validate the performance of ACA methods is essential when it is implemented in a specific application. For every case it is recommendable to measure the performance. In previous studies the nonparametric method has brought good results when it is applied for categorizing proportions (Ceron *et al.*, 2014).

2.1.5 Applications for Automated content analysis

Automated Content analysis is now also utilized to explore mental models, and their linguistic, affective, cognitive, social, cultural and historical significance. It is commonly used in tasks as pattern recognition, to identify and make predictions from data, identify and define concepts and topics (Nunez-Mir *et al.*, 2016).

In synthesis, the automated content analysis can be applied practically in any study that involves the analysis of large amounts of text, and where the content is important for the research such as opinions, speeches, transcripts, and any kind of recorded communication. It can be used in studies for marketing, propaganda detection, literature, cultural studies, sociology, gender and age issues, political science, psychology and many other related fields. The application in this thesis is determining sentiment of Internet users about climate change.

2.1.6 Climate change as a topic

Climate change is a topic of great interest for politicians and scientists as well as for the society in general. We can find information about climate change in social media networks such as Facebook, YouTube or Twitter; there is also information on internet in blogs, news or in scientific articles.

This topic is very controversial; it is difficult to find neutral content on the Internet. Experts on the subject, or research institutions have made most of the publications regarding the human impact of climate change. For the most part, these findings indicate that climate change occurs partly due to human activity. However, there is another group that although they accept climate change as a natural phenomenon, they claim that it has no direct relationship with human activity, they reject to take actions or change behavior to stop the climate change. As a result, the opinions of this group can discourage the daily actions carried out in society to stop climate change.

2.2 Machine learning

Machine Learning (ML) is described by Arthur Samuel as the field of study that gives computers the ability to learn without being explicitly programmed (Samuel, 1959). Machine Learning is programming computers to realize a given task using data previously collected. This input data is called Training Data and represents the experience, and this data is provided manually by humans or from large and sometimes complex data sets.

ML is a multidisciplinary field and can be related with artificial intelligence, probability, statistics, content analysis, information theory, as well as other important fields of science like philosophy, psychology or neurobiology. Sentiment analysis in Machine learning let us automate the process of analyzing large quantities of opinions from the social networks. With the use of ML algorithms, the nonparametric method reduces the human work. There exist different approaches to machine learning, the most common are supervised learning and unsupervised learning.

Some much used applications in ML are automatic translation, named entity recognition, speech recognition, classification and collaborative filtering (Alpaydın, 2014). We have seen that ML helps to automate most of the tasks of content analysis. There are multiple algorithms that are specially designed for the optimization of text analysis. Bayer Navies and Support Vector Machines are two widely used algorithms.

2.2.1 Supervised learning

Machine Learning techniques are used in sentiment analysis, which include methods of supervised and unsupervised learning. In this thesis the supervised approach will be applied to the nonparametric method. Supervised methods offer the advantage that it involves people in the process of labeling a small quantity of the total of document to study. The nonparametric method of Hopkins and King requires researchers to choose the questions and the data provide the answers. That is why the nonparametric method is considered a supervised learning method.

The scientist will oversee labeling the small sample choosing one of the categories chosen for the research in progress.

2.2.2 Other Machine Learning related fields

Probability and statistical theory are disciplines strongly related with content analysis and Machine Learning. Statistical theory is a discipline that collects, organizes and summarizes a large amount of data to generate relevant information relevant to the studied population. Statistics can be applied to almost any event, and therefore it is used in many scientific fields. Statistics are essential to determine the veracity of an event when there are cases of doubt (Cazau, 2006).

The proper interpretation of statistical methods, their input, and their results is the foundation of statistics. (Gooding-williams, 2017). In the literature of this thesis and in general in the field of Machine Learning, statistical terms are utilized, and these are necessary to understand. Statistics will also be an essential tool to validate the results of this research.

2.3 Sentiment analysis

Sentiment Analysis (SA) or opinion mining involve different fields like text mining, natural language processing, decision making and linguistics. SA is a type of text analysis that classifies, extracts and analyzes the opinions in the format of an unstructured text. The objective is to categorize opinions as positive or negative opinions associated with an topic involving people, organizations or social issues. Recently the objective of SA can also be the analysis of products and services (Singh and Dubey, 2014).

Sentiment analysis is a difficult task because of the complexity of the language and because the data is mixed (D. J. Hopkins and King, 2010). Automated methods will not replace the close and careful revision conducted by humans (Grimmer and Stewart, 2013).

2.3.1 Text mining

Text mining is a multi-disciplinary computer science field combining areas like information retrieval, data mining, Machine Learning, statistics, and computational linguistics (Ronen Feldman; James Sanger; 2007; Talib *et al.*, 2016). Text mining is also known as text analysis, which is the process of extracting structured data from unstructured blocks of text. The text mining literature is essential to understand the previous and complementary procedures of the method used in this thesis.

A library of texts will be analyzed with the Readme method, that data storage is also known as a corpus, which is generally stored as a text of strings. Computers normally manage the analysis of large corpuses; they perform a preprocessing of the text before the text is submitted for text mining.

The standard procedures for performing text analysis range from data preparation to analysis. That preprocess task mainly cleans the text of "stopwords" like numbers, punctuations, extra white space, word endings etc. keeping only the most relevant information for the further computational text management (Feldman and Sanger. 2007). Filtering and text preprocessing (converting to lowercase, removing punctuations and stemming) reduces complexity when natural language is converted to numerical variables. (D. J. Hopkins and King, 2010).

The next graphic shows the related fields, task and techniques involved with Text analysis:

Related fields task and techniques of text analysis

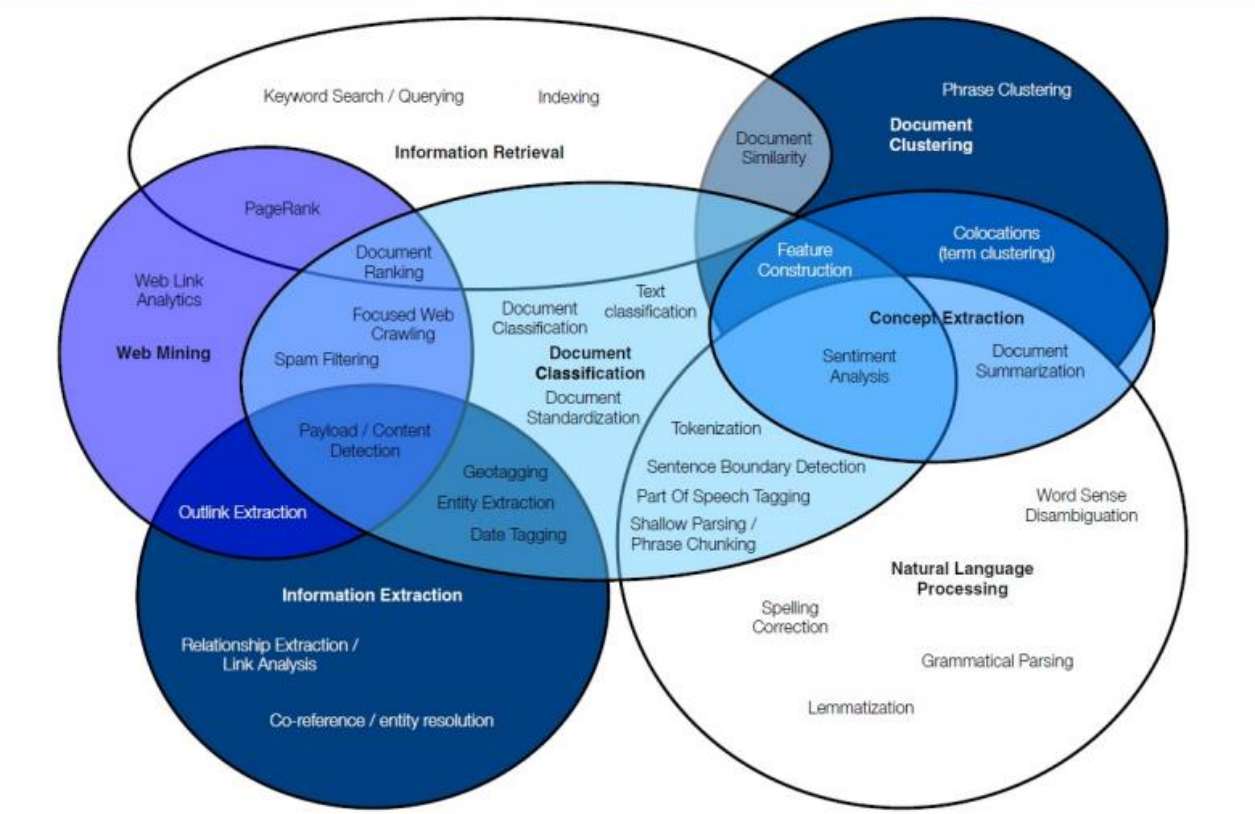


Figure 1 Inter-relationship among different text mining techniques and their core functionalities (Talib et al., 2016, p 415)

2.3.2 Text categorization

The basic cognitive process of arranging objects into categories is a fundamental process in human and machine intelligence and is central to investigations and research in cognitive science. Categorization is a key concept in the wide area of cognitive sciences, like for instance linguistics and philosophy (Henri Cohen and Claire Lefebvre, 2005). The process of categorization is a central task in any research. Until now, categorization has been approached from singular disciplinary perspectives with little overlap or communication between the disciplines involved. These disciplines could be for example linguistics, psychology, philosophy, neuroscience, computer science and/or cognitive anthropology.

The classification of texts facilitates the organization of information and defining the category of a text. There are methods that help to predefine categories according to certain characteristics present in texts. Sentiment categorization focuses on labeling the grade of opinion as for example positive or negative. Sentiment categorization is not as easy as topic classification (Pang, Lee and Vaithyanathan, 2002). One of the objectives of this thesis is to find the difficulties measuring opinions with social media content. I will make an exploratory study in the field of sentiment analysis.

2.3.3 Methods for text analysis

There are several techniques used in text analysis, such as individual, aggregated, supervised, and unsupervised. Individual models associate a specific topic with each data and tend to minimize the classification error for each tagged document. While they aggregate methods, they aim to estimate the final proportion of each category.

Some types of Automated Text Analysis (ATA) are Language detectors that automatically detects and tag documents according to a certain language. Sentiment Analysis (SA) is a type of ATA that identify the degree of positive or negative in texts containing sentiment or opinion. Summarization is another text analysis method that condenses long texts into consumable

portions. Classification is an ATA that classify and tags documents by topic, while Entity extraction is an ATA that extracts entities and values from texts.

In both individual and aggregate classification, a training set is needed. The training set could be manually or automatically encoded. The automatic tagging is usually based on dictionaries, or on the presence of certain positive or negative emotions. Methods such as Random Forest, methods based on decision tree, Naive Bayes, Maximum Entropy and Support Vector Machine are designed to analyze big data sets containing unstructured texts that put an emphasis on the social applications (Pang, Lee and Vaithyanathan, 2002; D. J. Hopkins and King, 2010).

Below I will show a general overview of the most known methods and how they are classified. Here we can see that the Hopkins and King method is measuring proportion and is in the supervised learning branch. On the other hand, the Bayern theorem is a method also supervised but it is in the area of individual classification.

The following diagram shows an overview of text analysis methods. In this way, we can appreciate that the method of HK is a task of classification where its categories are known and is supported by supervised methods where the objective is to measure the proportions of documents classified in categories.

A classification of automated methods for text analysis

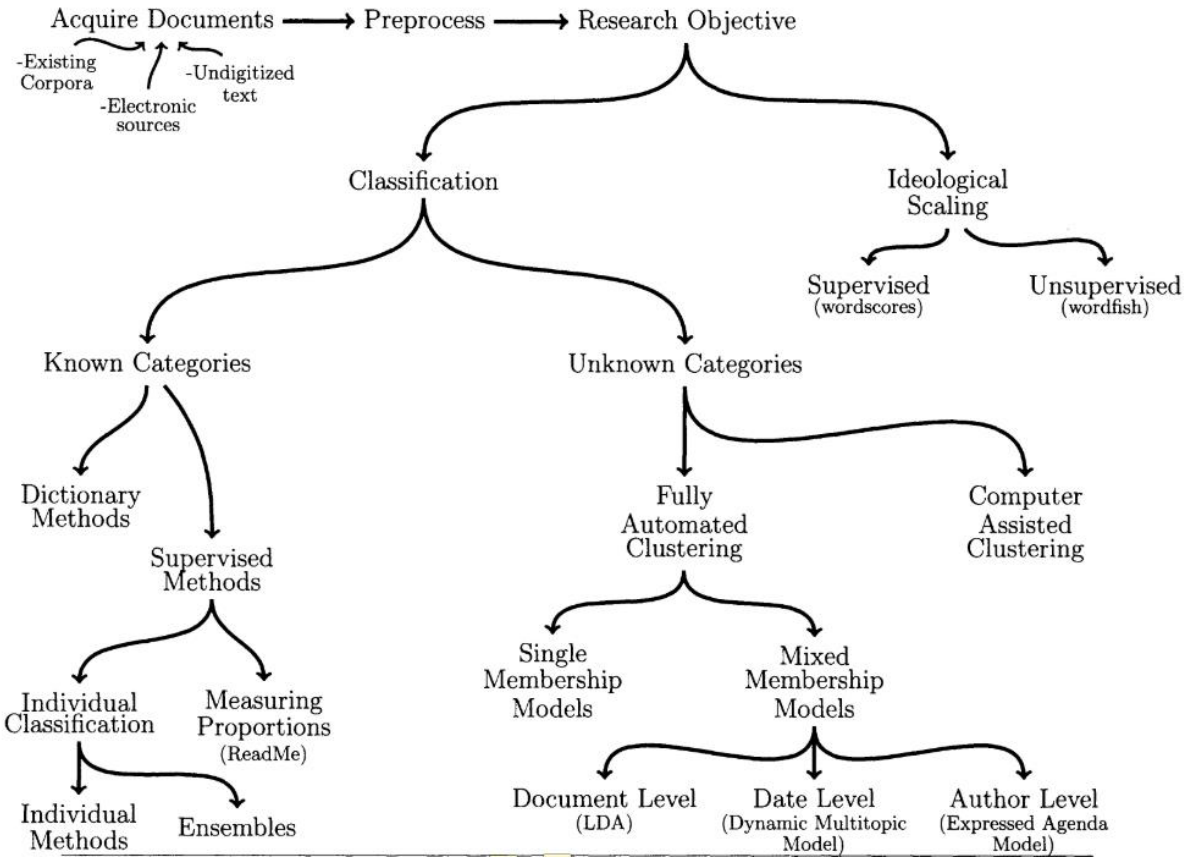


Figure 2: An overview of text as data methods (Grimmer and Stewart, 2013, p2)

2.4 The Nonparametric Method for Automated Content Analysis

This section brings a general introduction to the nonparametric method of content analysis (HK Method). It will be introduced with an overview of its origin and the authors. There will also be a description of the procedure and explanations of how it works. It will help to better understand the implementation of the method applied in the different applications.

2.4.1 The Nonparametric method

The nonparametric method for automated content analysis is designed to measure the proportion of documents in each given category, it is mainly used in social studies where large quantities of data is involved. The method has been proved with good results compared with other automated methods for text analysis (King and Lu, 2008; D. Hopkins and King, 2010a; Ceron, Curini and Iacus, 2014).

The method was developed by Garry King and Daniel Hopkins. Gary King is a Professor at Harvard University. He develops and applies empirical methods in many areas of social science research. He was listed as the most cited social scientist and has made some of the most important theoretical contributions in the field of automated content analysis (King, 2018). Daniel Hopkins is a Professor in the Political Science Department at the University of Pennsylvania. He is a political scientist whose research centers on American politics, with a special emphasis on racial and ethnic politics, local politics, political behavior, and research methods (Hopkins et al., 2018; King, 2018).

2.4.2 Advantages of the method

According to Hopkins and King the nonparametric method can be implemented for any study of the social sciences and can be used for researches in any language.

The method gives unbiased estimates proportions without the need of individual classification. That means that this method estimates the proportions directly. This is an advantage when using complementary classifiers with low accuracy. The categories are selected by the researcher, the subset must not be necessarily a random sample; It is also an advantage compared to the other methods, since in many social studies the random samples come from a different source than the population (D. J. Hopkins and King, 2010).

According to King, this method is the only nonparametric method developed for estimate multi-category proportions that does not resort to individual classification as a first step (King, 2018). Hopkins and King also state that this approach requires no modeling assumptions, no modeling choices and no complicated statistical approaches" (Hopkins and King, 2010).

2.4.3 Software Readme

King, Hopkins and Melendez designed the nonparametric method for automated content analysis as well as the software Readme. It is a package in R language that implements the method and it can run in Windows or Linux. The last version was released in April 2017. The Readme software takes as input a set of text documents, a categorization scheme chosen by the user and a small subset of text documents hand classified into multiple categories. "If used properly, ReadMe will report, normally within sampling error of the truth, the proportion of documents within each of the given categories among those not hand coded" (Hopkins et al., 2012 p1). The Readme software will be the tool that will be used to implement the method. Each one of the steps will be explained in the implementation chapter.

The Readme method can be complemented with other tools or methods of individual classification to facilitate tasks of classification of documents by a subject or to identify other

languages that are not relevant for the investigation. The main advantage of the Hopkins and King method is that even without these tools it maintains an accurate precision in order to estimate aggregate proportions.

3.4.4 The control file

The software ReadMe uses a control file in text format, where all the files are listed (labeled and unlabeled documents). This control file has three columns where the first columns contains the filename, followed by the category value and finally a binary value that indicates if the file is a training set document (value 1) or a unlabeled document (Value 0, or empty). All these columns are delimited by a comma.

A control file example looks like this example extracted from the ReadMe technical reference document:

Example of the Control file

```
filename,truth,trainingset
/users/m/readme/example/file1.txt,1,1
/users/m/readme/example/file2.txt,2,1
/users/m/readme/example/file3.txt,2,1
/users/m/readme/example/file4.txt,3,1
/users/m/readme/example/file5.txt,,
```

Figure 3: Example of the control file in the corpus. (Hopkins et al., 2012)

2.4.5 The training set

The training set is represented by a subset of documents of the total population. This is a small sample (sometimes random) drawn from the total document population or another related corpus. It is represented by "I" where it obtains values from 1 to n where n represents the total of documents in this small random sample. All these documents will be labeled with one of the categories selected by the investigator.

The training set does not necessarily need to be a random sample. This is a great advantage compared with other methods. The minimal quantity of documents needed to be classified are 100 documents (D. J. Hopkins and King, 2010).

According to the Hopkins and King literature, coding as few as 100 documents is enough for most applications. This represent an advantage when we want to choose a method that let us reduce excessive cost in hand coding and time consumption.

"Coding more than 500 documents to estimate a specific quantity of interest is probably not necessary, unless one is interested in much more narrow confidence intervals that is common or in specific categories that happen to be rare. For some applications, as few as 100 documents may even be sufficient" (Sterne, 2010 p.99). The aggregate proportion is represented by ($i = 1, \dots, n$)

2.4.6 The subset

The unlabeled documents, also known as inferential target, is a large set of documents that will be processed by the computer. We can present these documents as l ($for\ l = 1, \dots, L$) With an unobserved classification Di (Document category variable)

All these documents are included in the corpus, together with a control file that contains the list of documents with the labeled category (the training set).

2.5 Related Works

This section covers other related works in order to compare their results with the method applied for climate change. This section explains the running example used by Hopkins and King to explain the method. This section is included because I want to confirm that this method works with different social applications.

2.5.1 Verbal autopsy

This method has its origins in the work of Garry King and Ying Lu in 2008, "Verbal Autopsy Methods with Multiple Causes of Death", where it is intended to solve the problem of estimating causes of death through a method that allows the classification into multiple categories. (King and Lu, 2008).

The verbal autopsy is a practice used to analyze the information provided by the caregivers about the symptoms observed before death. This standard procedure calculates proportions of document categories mainly used in undeveloped countries in regions of Asia and Africa, this practice is commonly used for estimating the cause of death. Parametric methods analyze only one mortality cause at the time, making the procedure expensive, time consuming or unreliable (Soleman, Chandramohan and Shibuya, 2006; King and Lu, 2008; World Health Organization (WHO), 2012).

"Current approaches can analyze only one cause at a time, involve assumptions judged difficult or impossible to satisfy, and require expensive, time-consuming, or unreliable physician reviews, expert algorithms, or parametric statistical models."(King and Lu, 2008 p. 78) The method was successfully implemented to face this problem. (King and Lu, 2008). This is a practical example where the nonparametric method of HK is required.

2.5.2 Measuring opinion about the US election in 2008

In Hoping and Kings (2010) article about measuring opinions in the 2008 US election, the Readme software was used. This demo application shows the process to prepare the data and the methodology used to collect the information. The public opinions about the American presidency was measured with the HK Method. The opinions about President Bush and the 2008 candidates were analyzed from more than 10,000 blog posts focused on President George Bush using keywords like "Bush," "George W.," "Dubya," or "King George" and similarly for the rest of the candidates. 442 post were hand coded by researches into the categories (−2) extremely negative, (−1) negative, (0) neutral, (1) positive, (2) extremely positive, (NA) no opinion expressed, (NB) not a blog.

When the method was applied, it reveled changes in the public opinion about John Kerry after he said, "You know, education—if you make the most of it ... you can do well. If you don't, you get stuck in Iraq" where the public opinion became extremely negative after that joke (Hopkins and King, 2010, p. 231). In the same study, the HK method is implemented in other applications. The study concludes with good results when the method of HK is applied. It shows high accuracy estimating category proportions compared with other supervised methods for text classification. The HK method shows good results even when the labeled set is in the range of 100 to 300 hand labeled documents analyzing large quantity of documents.

2.5.3 Measuring the 2012 elections in France and Italy

In 2012 the Readme method was applied to find the citizens political preferences in France and Italy (Ceron et al., 2014). Online popularity of Italian political leaders and the voting intentions in France was tracked, both in the 2012 presidential election and in the subsequent legislative election. Traditional offline surveys were also monitored, and then compared with the actual electoral results.

This study included Twitter as a data source. This social media platform has increased their number of users steadily in the last couple of years. Twitter is widely used to issue opinions

and is an important source because it is widely used by all kinds of influential people and politicians. (Hambrick *et al.*, 2010). In the article of "Every tweet counts?" Ceron mentions the advantages by working with social media compared to traditional surveys, as well as mentioning the advantages of the Hopkins and King method compared with other traditional sentiment analysis techniques (Ceron *et al.*, 2014). The study concludes that the method of Hopkins and King produce more accurate results compared with traditional surveys.

According with the mentioned related works, exists immense possibilities to use this method in many social science applications, because it is possible to analyze every type of unstructured text. Additionally, there are a lot of sources of digitalized data everywhere; blogs, emails, articles, digitalized books, etc. This makes the method a very flexible tool.

3. Methodology

Around 40 videos with high numbers of visualizations were selected, 20 videos from “official sources” and other 20 from “unofficial” sources according with the climate change movement. All these opinions were extracted on May 20th of 2019, most of the videos were published between 2010 to 2019.

The focus of this thesis is to apply the method of Hopkins and King in order to know if it is a reliable method for measuring the percentage of opinions of users of social networks with respect to climate change. This study may serve as a preliminary study in opinion mining research applied in any topic. The result of this thesis will indicate if this method is reliable for this specific topic, climate change.

It is worth mentioning that the main objective of this research is to measure the performance of the Hopkins and King method, so the information selection methodology regarding climate change would not be very relevant, the main objective of this thesis is to measure the performance of the Hopkins and King method, so it can be considered for a further research for the topic of climate change.

The videos were selected through a quick search by relevance, and then manually classifieds by type of content (from activists or deniers). In total, 40 videos were analyzed and their comments extracted, 20 videos were made/posted by activists and deniers posted the following 20 videos. Around 901 extracted comments were manually categorized in values from -1 to 4, {-1 for deniers, 0 for neutral viewers, 1 for activists, 3 for videos no relevant and 4 for comments difficult to classify for humans}. Next, the methodology for this study is discussed.

3.1 Tools and software

Automated tools for sentiment analysis are indispensable when large quantity of data must to be analyzed. For this research the use of technologies as web crawlers, or scrapers are mandatory to extract opinions from social media.

I have used the tool “YouTube comment Scraper” (Klostermann, 2015) to extract the comments and other relevant information as quantity of likes / unlikes, the user name, date of the post, etc.
I searched from YouTube manually by using the keywords "Climate change" or "Global warming". The output from “YouTube Comment Scrapper” are a list of documents in csv format.

According to the literature of the nonparametric method, it is possible to use conventional classifiers to avoid human effort and time consumption, although it is not mandatory, even though there is no problem if the classification accuracy is low. This is one of the advantages using this method. Even without the filtering of opinions, it can give good results (D. J. Hopkins and King, 2010).

The R project for Statistical computer(*The R Project for Statistical Computing*, 2018), is a free software environment for statistical computing and graphics. This tool is widely used for text mining, and this thesis is mainly implemented in R language. There is a wide collection of R packages to use for data mining and sentiment analysis. The method of Hopkins and King is implemented in the package “ReadMe” and “VA”. I also used another R packages as “quanteda”, “NLP”, “tm”, “caret”, a full list of tools is listed in the appendix tools.

The R package “ReadMe” implements the HK Method. The operative system Linux is recommended to run the ReadMe Software. I found many difficulties and issues regarding the software while I run it on Windows. It is recommended also to install Python before the installation of R. RStudio is optional to use, RStudio let users have a friendly user interphase. The complete list of tools is showed below.

List of tools to run the software ReadME

1. Linux (recommended) /Windows / IOS
2. Python 2.7
3. R 3.4
4. Devtools Package
5. RStudio (Optional)
6. Library VA
7. Library ReadME

In the appendix section one will find a list of the tools mentioned in the thesis, as well as software and packages used in this implementation.

3.2 The target audience

Each publication on the Internet is intended to inform, entertain or express an opinion regarding a specific topic. In this study I analyzed manually around 900 opinions issued by Internet users within the YouTube video platform. I found that videos are an important source to extract opinions. In a video for example, the effects of the climate change are visually displayed, causing a considered impact to the Internet users.

As the objective of this thesis is to detect those who believe that climate change is caused by human activity and those who hold the opposite believe, the profiles of the users were analyzed, which in turn will be classified into one of the following categories:

The activists. Will be all those who show a favorable opinion regarding the content that is being shown, given that the content is a video in favor of human made climate change based on clear evidence or serious investigations. It will also be those who voice their opinions in a series of statements to help to stop climate change, for example, to stop using the car, to consume less beef, or encourage the use of clean energy.

Some examples of comments made by activists:

File	Comment
A01-31.txt	"You just have to look at extreme weather events happening more and more frequently."
A01-41.txt	"I can't believe that some people are convinced that it's all a hoax. I really hope they're right, but reports like this scare me to death."
A01-45.txt	"Trump is an idiot "
A01-103.txt	"And yet the climate denier assholes still insist nothing is wrong"

Table 1: Examples of comments made by activists

The deniers. Are all of those who demonstrate a clear rejection of scientific evidence regarding human impact on climate change. There are those who consider that this change is due to natural cycles that the earth has had throughout its history. These users clearly show a refusal to take actions in regard to reduce the human impact on the environment, relying on ideologies, policies, customs or religion.

Some examples of comments made by deniers:

File	Comment
A01-72.txt	"Biggest fraud Science."
A01-109.txt	"In 1100-1300 it was almost as warm as it is today."
A01-203.txt	"Buy a coat, we are entering another solar minimum, Maunder Minimum. Figure out to stop an ice age. We are in a mini warming between ice ages. These warmer temperatures allowed us civilization. "
A01-212.txt	"Don't let these scam artists upset you. This is all bullshit to get you to pay more taxes and give up more freedom. Watch Tony Heller on YouTube. He pulls back the curtain on junk science."

Table 2: Examples of comments made by deniers

The observers. Their opinion is not very relevant, they are integrated into the conversation, but they do not issue a clear opinion about it. Sometimes their subject is not related to the discussion or they can even post propaganda that can be considered spam.

After I have hand coded some opinions, it is easy to identify if a comment is made by activists or for deniers, from the type of vocabulary used. Excluding the most frequent words used for both groups ("climate", "change", "people", "like", "global", "warming", "just", "can", "years", "Need") the general conversation of the activists and deniers is visualized as follow:

The difference in the conversation between deniers and activists:

Activists wordcloud

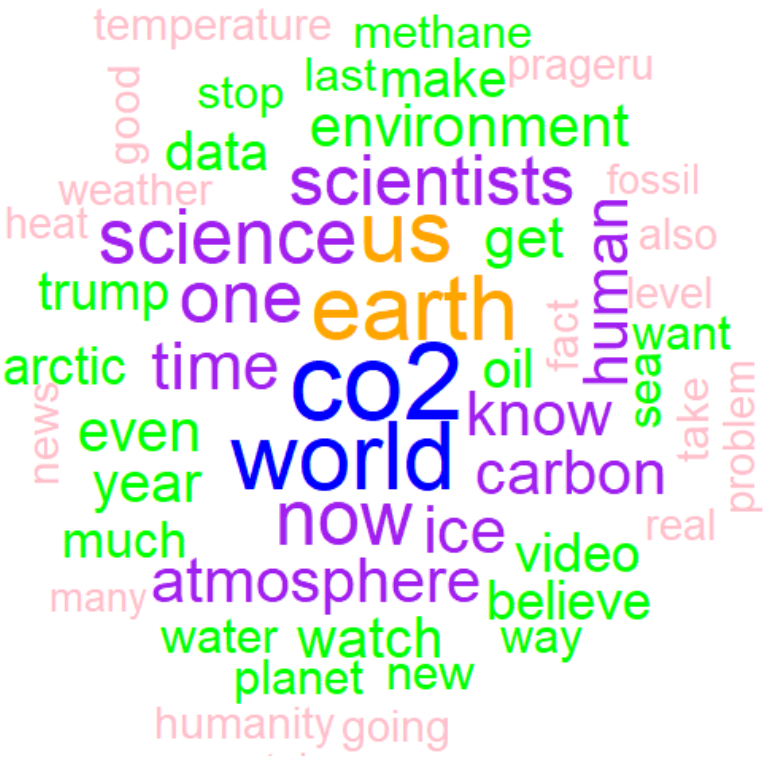


Fig activists wordcloud

Deniers wordcloud

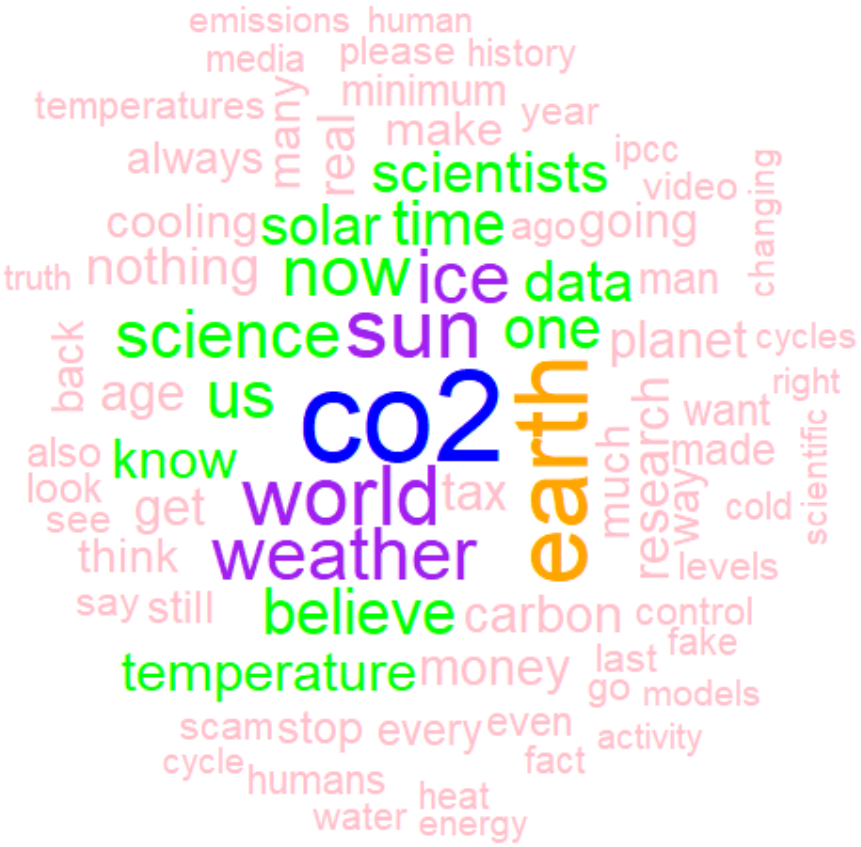


Fig. Deniers wordcloud

We can observe the difference in the words used for every group, while the activists talks more about the reduction of the ice in the Artic, Donald Trump, the environment, oil or atmosphere, the deniers talks about the minimum solar, tax, ice age, fake news, cycles or control, to mention some of the most frequently words used. Empirically it is possible to estimate the category of an opinion with the presence of some keywords.

3.3 Collecting the data

The data collection mechanism is optional, is it possible to choose any technology for this task, and there exists many different tools and packages. To setup the computer, it is recommendable to follow the steps described and use the packages and tools implemented.

For the research, two types of data collections was selected; the comments from videos posted from activists and organizations and the comments from the videos posted from deniers. In total 20 videos posted by activists or organizations and other 20 videos published from deniers were analyzed.

For the videos from activists and organizations, I searched on YouTube using the keywords "Climate change" and "Global warming". From the list I choose the videos created by recognized organizations, politicians and activists that post the videos with the purpose of inform the population about the consequences of climate change. I have seen and verified every video to be sure the content is in favor to change the human behavior to reduce the effects of the climate change. For the videos from deniers, I used the keywords "climate change" and "Global warming" adding other keywords like "hoax", "myths", "skeptic", "lie" and "deniers". The full list of videos analyzed is attached in the appendix D.

After I collected and documented all the links of interest, the next step was to extract the comments using "YouTube comment scraper". This is a web-based tool that extracts comments from YouTube. It is free and open source licensed under ISC. (Klostermann, 2015). As output "YouTube comment Scraper" store the comments on documents in CSV format. This tool also collects other important information as published date, duration, total views, likes, dislikes, and number of comments. The complete list of the information of the videos is in the appendix E. "General stats of the analyzed videos"

3.4 The data source and document identification

The documents has been listed with an specific format that contemplate the name of the file, the classification of content and a unique id number

[source][video number]-[opinion number].[document format]

Source: to identify the source (a: for activists' source, b: for deniers' source)
Video number: The number of the video, useful to identify the video in question.
Opinion number: The document number used as identification and unique reference.
Document format: The format of the file, for this study the txt format is used.

Examples of file names:

File name	Description
a1-203.txt	Represent a text document with the opinion nr 203 extracted from the video nr 1 published by activists.
b2-35361.txt	Represent a text documents with the opinion nr 35361 extracted from the video 2 published by deniers

Table 3: The representation for the standardization of the corpus

As I mentioned previously, the videos themselves are a kind of opinions that must be categorized as the first level in activists or deniers. Negative opinions in both kind of videos will produce biases, because negative opinions in pro climate change videos are opinions that will be categorized as denier, the same text could appear in videos anti climate change, that must be categorized as opinions from activists.

In the following, two graphics are presented. The first one shows the list of activists' videos with the percent of the likes and dislikes. The next graphic shows the deniers' videos and their percentages of likes.

Percent of likes and dislikes in activists' videos

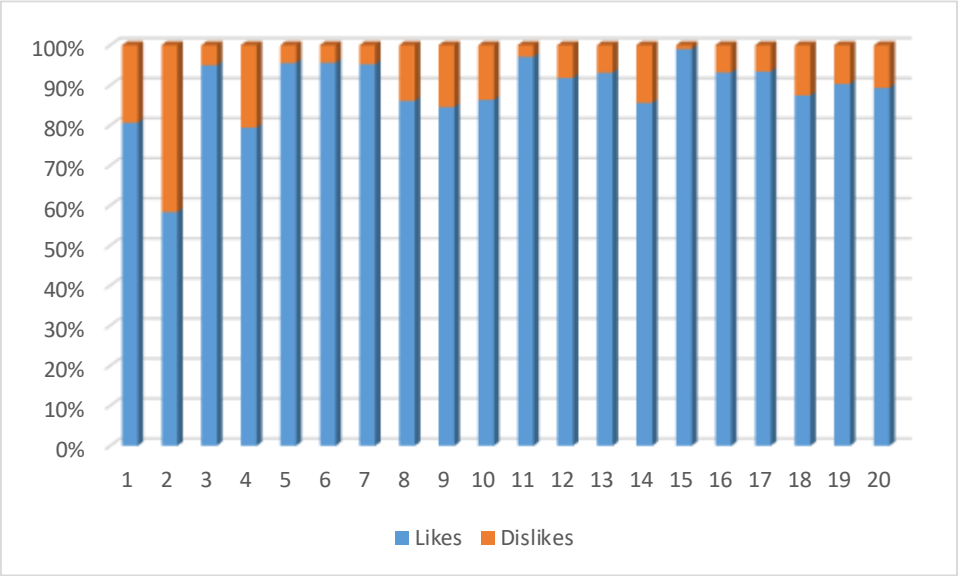


Figure 4:Shows the percent of likes and dislikes from videos published by activists.

On the previous graphic we can observe a clear tendency in likes regarding the video from activists. These numbers denote that the viewers agree with the content of the video. But what happens with the number of comments related with each video? We still do not have much information regarding if the comments are positive or negative according with the content, or if it is possible to detect if the opinions comes from activists or from deniers. In the following, I will analyze the proportion of likes and dislikes from the deniers.

Percent of likes and dislikes in deniers' videos

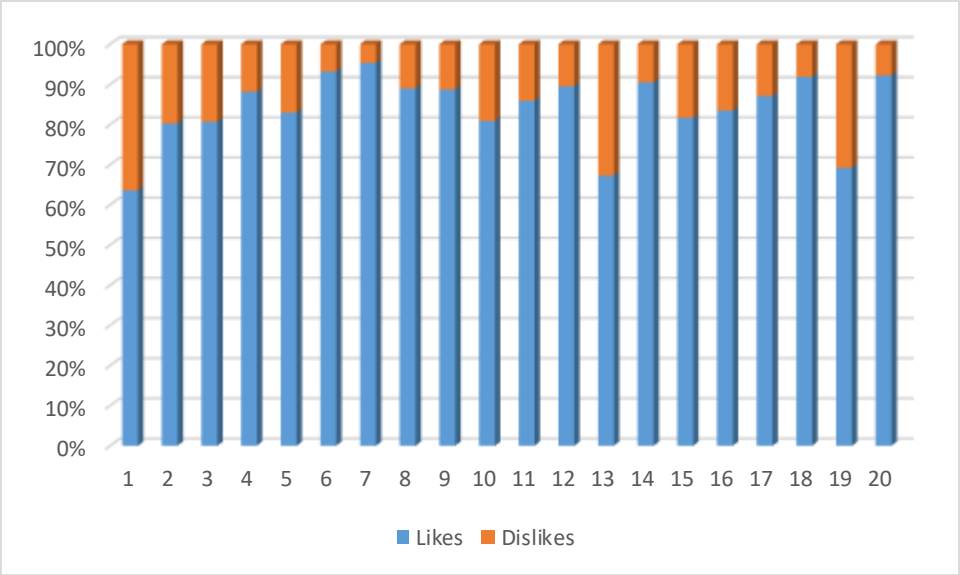


Figure 5: Shows the percent of likes and dislikes from videos published by deniers.

In figure 5, we can observe the same tendency regarding likes in the videos from deniers' sources as well. It shows that there are two segments of audience that must be analyzed in separated studies. After getting an overview of the proportions of likes and dislikes. In the section 5, I will continue applying the HK Method to measure the proportions of the opinions in each category (for deniers or activists).

3.5 Category selection and coding

The method in question has the ability to work with multiple categories, which will be chosen in order to classify the extract of the document according to the degree of opinion. These categories must be mutually exclusive and exhaustive, that means, from the chosen categories it is only possible to select one category for each document proportion, excluding the rest of the other labels. These labels will also be represented by a numerical value that will determine the degree of opinion.

The document category is composed for the selected grade of opinion and other labels that helps to exclude other possible values that do not contain a grade of opinion. The document category variable is represented for Di where Di = j for categories j = 1, … J. As example Di can take values of {-2, -1. 0, 1, 2, NB, NA}

Categories elected in the Hopkins and king Study in 2008

Value	Category
-2	Extremely Negative
-1	Negative
0	Neutral
1	Positive
2	Extremely positive
NB	No opinion
NA	Not available

Table 4: categories used for measure the opinion about candidates of the 2008 US Elections (D. J. Hopkins and King, 2010)

As can be seen in the table above, the categories are represented by numerical values for later analysis, while the NA and NB values will be considered to make the classification mutually exhaustive and exclusive. In this specific case, using blogs or comments from social media like YouTube, it will be difficult to find the NA because blogs usually express an opinion.

The total of the videos was fully watched, and then classified into content from activists or deniers. Then the opinions were extracted from the videos and creating a document for every opinion, the documents was named with a format designed to easily identify and group the documents for this study.

A total of the 601 opinions from the video a-01 were stored in a separated document, each document contains an opinion. All of them were labeled manually and stored in the control file control_a01. For the first experiment 600 documents will be selected to test the method with different subsets and validate against known categories. The entire subset of labeled documents brings flexibility to experiment in different scenarios.

Before labeling the documents, it was important to recognize the profiles of the target audience as showed in the section 3.2. I read and labeled the selected documents following the instructions:

“Classify the followed opinions related with the Climate change, choose -1 if the opinions are made by deniers; 0 for neutral; 1 if the opinions are made by activists; Choose 3 for all the comments that are not relevant for this conversation; and finally choose the option 4, if the opinion is difficult to understand or to identify if the comment is positive or negative to the fact that climate change is affected by human behavior”

How would you classify the following opinion?

The coder has the following options to choose from:

Category	Label	Description
Denier	-1	The opinion has the propose of refute scientific evidence about the climate change.
Neutral	0	This people emit a neutral opinion
Activist	1	The opinion is done with the propose of aware another people about the effects or to take actions to stop the climate change
Non relevant	3	If it is not related to the subject. It is spam, or any other answer that does not have to do with the previous ones.
Non understandable	4	If this kind of text has a tended opinion but it is difficult to determine if comes from a denier or from an activist. Sometimes is a kind of sarcasm difficult to classify, this kind of opinions are no possible to determine. This is a hard task for the human and the machine as well.

Table 5:List of categories for selected for this thesis

The coding instructions are useful in studies with more than two coders are participating, it is important to provide a previous training to avoid biased in results because a non-standardized procedure for labeling. These procedures involve training for coders, evaluation by analyzing inter-coder reliability rates, and getting feedback from the coders. (Melendez *et al.*, 2018)

For this work I skipped the use of multiple coders to avoid biases in the results and to avoid excess of work. This approach doesn't affect the objective of the study of the HK method on this thesis.

3.6 Data Filtering

Standards procedures for filtering are classify the documents individually through traditional automatic classifiers, such as Support Vector Machines, Naive Bayes etc. This first phase will help select the documents that are closely related to the subject. Filtering the documents will save time and effort, obtaining only the opinions that are relevant for the research. This step is not necessary since the method can perform well without this step, but it will be a way to save time.

As previously mentioned, the videos were classified into deniers and activists. This kind of filtering has not used any automated tool. This study does not need a large quantity of videos to test the performance of the HK method, but for large studies it is possible to use automated tools for classification. Every video on YouTube has a short description in text format; it can be scraped and classified using any kind of technology.

It was easy to identify videos published by activists because they are made by professionals. Here I include scientists, organizations, news agencies, etc. These videos are characterized by the fact that they are based on serious investigations, where the aim is to inform in an objective manner. On the other side are some recognized deniers that publish periodically content about climate change or global warming with keywords like “hoax”, “Minimum solar”, “lie” or “fake news”.

As part of the methodology I have hierarchized the grade of opinion into three levels. Analyzing the YouTube videos according with their opinion, the first level categorizes the video. Every video with political content is published to express opinion about a specific topic. On this first level it is necessary to categorize by sentiment or grade of opinion to avoid biases. On the second level the direct opinions from the users of the YouTube platform are placed and in the third level are the comments of other opinions or responses of the comments. For this study the opinions in the third level are omitted.

There was not implemented any other special filtering after the opinions in the third level was dropped. In this way the method is tested when exhaustive categories are used as filter to compare with the same data previously filtered with other tools or methods.

I am using the category 4 (opinion non-understandable) as exhaustive category. It must be considered that within activists or deniers, the poor quality of language or even the advanced use of sarcasm could confuse the coder and even more the computer. For this type of opinion, the use of exhaustive categories is essential. In practice this special category "4" has to works as a filter for as non-understandable opinions.

The method works with any language but in many applications the language is limited to one language in order to avoid complexity and reducing time and effort in the human coding process. In most of the applications, English language is used the most, so a basic first step is to drop non-English language. For this, any kind of individual classificatory that detects languages can be used.

"Our method works without filtering (and in foreign languages), but filters help focus the limited time of human coders on the categories of interest" (Hopkins and King, 2010, p 232).

3.7 Text preprocessing

In order to facilitate the implementation, it is necessary to manipulate the unstructured text to make it more comprehensible for the computer. “Preprocess the text within each document by converting to lowercase, removing all punctuation, and stemming” (Hopkins and King, 2010, p. 232). When preprocessing texts, we can reduce its complexity and eliminate irrelevant information.

In this step, unsupervised methods or manual methods to process the text can be used. It consists of converting all words to lowercase, remove scores and keep only the root words using N-Gram-based techniques. For example for the word *consistency* there are derivatives such as consistent, consist, consist etc. (Cavnar, Trenkle and Mi, 1994).

For text processing, the steps suggested in the HK literature will be followed. The preprocessing of text could be performed, using unsupervised methods. To test the method in different scenarios and to find the best performance for these kinds of studies (YouTube + Climate change) a combination of the followed preprocessing task will be implemented. The objective of these activities is to find the best text preprocessing combination or simply find the preprocessing that are not necessary to skip their implementation and save resources for large studies.

The tasks of tokenization and bag of words are standards preprocessing text activities in opinion mining and mandatory for the implementation of the HK method. For the preprocess of stop words / threefold I will experiment with different ranges in Threefold [0.00 : 0.10] and stop-words to compare the results between them.

For the tasks of to lower case and remove special characters I will just follow the recommended standard text preprocessing that consists in converting to lowercase all the documents and remove punctuations, digits and special characters before running the HK method. In relation with n-grams, this thesis only experiment with unigrams

For the task of “Concatenate the username”, I found that it was easier to classify opinions when I read other opinions from the same user. It helps me to label some comments that were not so easy to classify. It is probable an interesting experiment for a future study.

The tasks to be performed in the data preprocessing are:

Task	**Use in HK**	**On this thesis**
Tokenization	Required	Implemented
Bag of words	Required	Implemented
To lower case	Optional	Standard
Stop Words / Threefold	Optional	Implemented
Remove punctations, digits or special characters	Optional	Standard
Stemming or convert to n-grams	Optional	Only unigrams
Concatenate the username	Optional	Future work

Table 6:List of preprocessed task used in this thesis.

5. The implementation with Climate change as a topic

In this section the practice of the HK method will be applied on the topic of climate change. The steps needed are described in details with their code in R language. The implementation of the method is divided into two stages. The first part of the method will represent the text in numbers. The second part is to code variables for a posterior analysis and validation. For the first part, it is necessary to perform some previous steps such as filtering the documents and preprocessing the text. These steps will be elaborated further in the text.

Tasks such as collecting information, filtering and preprocessing of the text are the first steps to prepare the text before applying the HK method to calculate our quantity of interest. Unsupervised methods and tools are used to realize tasks such as web crawling, information retrieval, topic classification, language detection, etc., and all these tasks help to reduce the human effort.

In the second phase the hand coding is indispensable in order to classify the subset of documents. These tasks are also explained in this chapter and applied in the chapter of implementation.

5.1 The Procedure

To implement the HK method it is necessary to perform some complementary tasks that can be mixed with other methods or tools. Most of these tasks are not mandatory and gives to the researcher the freedom to choose according with their methodology designed. This is an advantage in implementing the HK method according to the author.

In the first step the videos were non-randomly selected from the platform YouTube. The data was filtered manually and the videos were classified into deniers and activists. I proceeded to hand code the documents in the "control file". In this step it is necessary to experiment with

different preprocessed data in order to later preprocess the text in multiple variants. Finally, the function readme is implemented, and the results are validated via bootstrapping and cross validation.

The steps for the implementation:

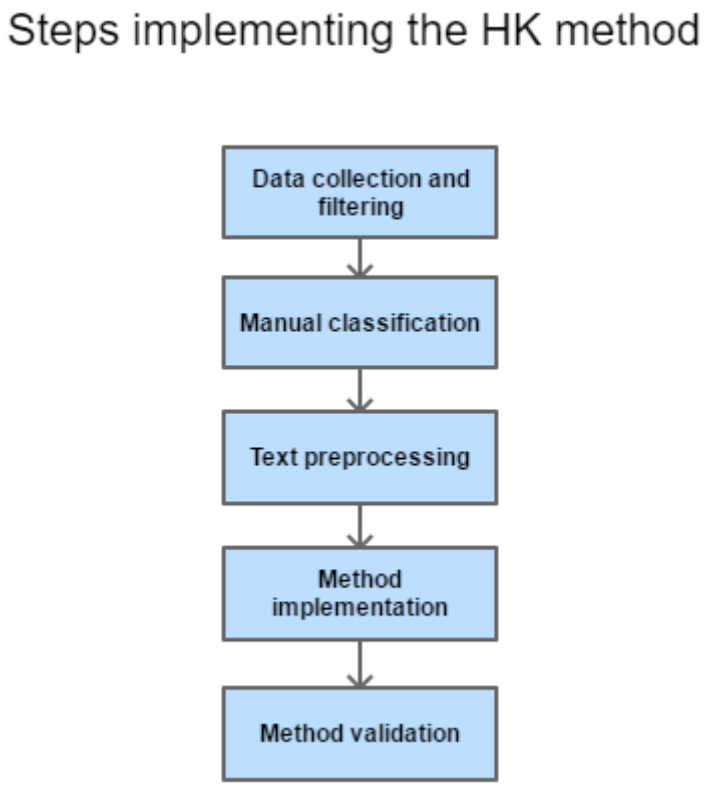


Figure 7 Implementation of the HK Method

5.2 Set up the environment

One of the first steps is to create a folder corpus under the ReadME library to store the csv files. Commonly the R packages are installed in "/usr/lib/R/library". The data collected for this thesis can be download from (Martinez, 2019), and unzipped in the folder corpus. All the files can be stored in "~/R/library/ReadMe/corpus/csv_files/all".

Following the methodology for this research, I have scraped the opinions and stored in a folder (inside the R library ReadMe) with 40 csv documents. Every document has been imported to R into a data set. With a script I drop the information that is not relevant, keeping the opinions in the first grade and some other relevant information as id file.

With this script I can obtain the full list of opinions:

```
#Set the working directory and list the csv files inside
setwd(system.file("climate_change/csv_files/All", package="ReadMe"))
files <- list.files(pattern="*.csv")

#Store all the opnions
opinions_csv <- do.call(rbind, lapply(files, function(opinions) read.csv(opinions, stringsAsFactors =
FALSE)))

# Keeps only the columns containing relevant information from the dataframe
opinions <- opinions_csv[c(1:7)]

#Delete empty rows
opinions <- opinions[!apply(opinions[c(2)] == "", 1, all),]
```

Script 1: Setup the enviroment

It generates a dataset with the full list of opinions:

Dataset with the full list of opinions

	file	id	user	date	timestamp	commentText	likes
1	a01	Ugwb5IF5jy-KXRX49CB4AaABAg	Lucy Balls	1 hour ago	1.56e+12	We have been so dumb I wish we could start over	0
2	a01	UgyIZIGRfstel_aiyo14AaABAg	Raul Ortiz	3 days ago	1.56e+12	The jet engine is the epicenter of all climate change. There i...	2
4	a01	UgzBKJdSshvsR7sLH1N4AaABAg	Renee Tyler	3 days ago	1.56e+12	People are so dumb today they're like sheep going into Slau...	1
5	a01	Ugx5jjlJTUK-kss7cZJ4AaABAg	WE OBEY JESUS	3 days ago	1.56e+12	I think they say it is cooling now. It's not science.	0
6	a01	Ugz9ZzwPo7IzcOvdz-J4AaABAg	Mia silva	3 days ago	1.56e+12	THE TRUE IS, WE DO NOT HAVE EVEN 5 YEARS ...ITS GOING...	0
7	a01	UgxXGzf2ccqmCWHd-aJ4AaABAg	Mia silva	3 days ago	1.56e+12	I think president Trump doesn't want to say he doesn't belie...	0
8	a01	Ugw_ybUWRN1L93zOEbR4AaABAg	geoff mcintosh	6 days ago	1.56e+12	work in the pilbara desert for the last 10 years.Â cooler now...	0
9	a01	UgyZE7QaOJKCoNgleJR4AaABAg	geoff mcintosh	6 days ago	1.56e+12	happy to buy any coastal properties for $50. Doing you a fa...	0
10	a01	UgzUHUkgVP068G5DLGV4AaABAg	hamish counsell	1 week ago	1.56e+12	This is where i beg the question, why are we still not looking...	0
11	a01	Ugwo_PX2w1HcdGiRcbJ4AaABAg	Thomas Barry	1 week ago	1.56e+12	GODS HAND NOTHING YOU CAN DO EXCEPT REPENT AND...	0
12	a01	Ugw2mQ3OzRRsdtN05qt4AaABAg	John Kane	1 week ago	1.56e+12	Climate change is real all people care about is their money a...	0
13	a01	UgyUK01m7-o6rDGXl3t4AaABAg	Alex N	1 week ago	1.56e+12	13 of the last 16 years have been the warmest years on reco...	2
14	a01	Ugz1LcqmR5OEJQcALFB4AaABAg	Jai Singh	2 weeks ago	1.56e+12	Just like they made us believe the Y2K bug but that after bill...	0
16	a01	UgzkjjdEJVk-mBWU-k84AaABAg	Father Andrew	2 weeks ago	1.56e+12	Complete and utter propaganda! Those scientists who inter...	1
18	a01	Ugx0cJapp8JayuU2vpt4AaABAg	ZomBie guy MansTer	2 weeks ago	1.56e+12	This is just the beginning im serious this is the kalyug im a h...	0
20	a01	UgzuoWtxaNO8DNludcd4AaABAg	Lost With Lewi	2 weeks ago	1.56e+12	Bring on the heat baby ! I love hot summers	0
21	a01	UgyDxhHraPFCTmmHZIF4AaABAg	noushin saeedi	2 weeks ago	1.56e+12	People...!!! HARRP.....Look it up.	0
22	a01	UgxLAP8x3N7ym3W_QAx4AaABAg	Ray Eason	3 weeks ago	1.56e+12	Why do I keep finding videos featuring credible scientist po...	0

Figure 6: List of the dataset containing the opinions before the manual labeling

Building the corpus

In order to apply the method it is necessary to create a corpus containing the large subset of documents to analyze, every document contains a comment from the dataset imported. The document must be created in csv or text format. All these documents will be preprocessed and analyzed by the readme software.

Generation of the corpus:

```
#create corpus
#----------------------------------------------------------------------
x_opinions <- opinions[c(6)]
n_opinions <- unlist(x_opinions, use.names = FALSE)
typeof(n_opinions)

for (i in 1:length(n_opinions)) {
  write.table(opinions[i,c(6)], file = paste0("E:/R/R-3.6.0/library/ReadMe/climate_change/corpus/00/",
opinions[i,c(1)] , "-" , i , ".txt"), col.names = FALSE, row.names = FALSE)
}
```

Script 2: Building the corpus

A corpus containing 83,146 text documents have been generated in the folder 00.

The ReadMe software needs a control file to specify the list of documents to be used as the subset and the files that are used as training set. The control file (control.txt) specifies the files names, the type of document (Training set or test set) and the value assigned representing the document category. These 3 columns could be separated by coma or space.

The control file

ROWID	TRUTH	TRAININGSET
a1-345.txt	1	1
a1-355.txt	1	1
a1-387.txt	0	0
a1-459.txt	-1	0

Figure 7: A simple representation of the Control file

In the column ROWID are the names of the files that will be analyzed, the TRUTH column contains a value between -2 to 4, representing the categories for this study, and finally the TRAININGSET column that contains a binary value to specify if the file is a training set (value 1) or a test set (value 0) .

5.3 Manual Classification

This step is one of the most important activities of this research and the most time consuming. It is important to take time to read carefully every opinion and find the easiest way to read every opinion and capture the category in a numeric value. I will process to export the "opinions" dataset into an excel file. I felt comfortable doing the categorization this way, as it brings me flexibility to read the documents at the same time as I perform the task of labeling.

This step is optional, it is also possible to hand code the category directly to the control file. With the following script the excel file for labeling is created.

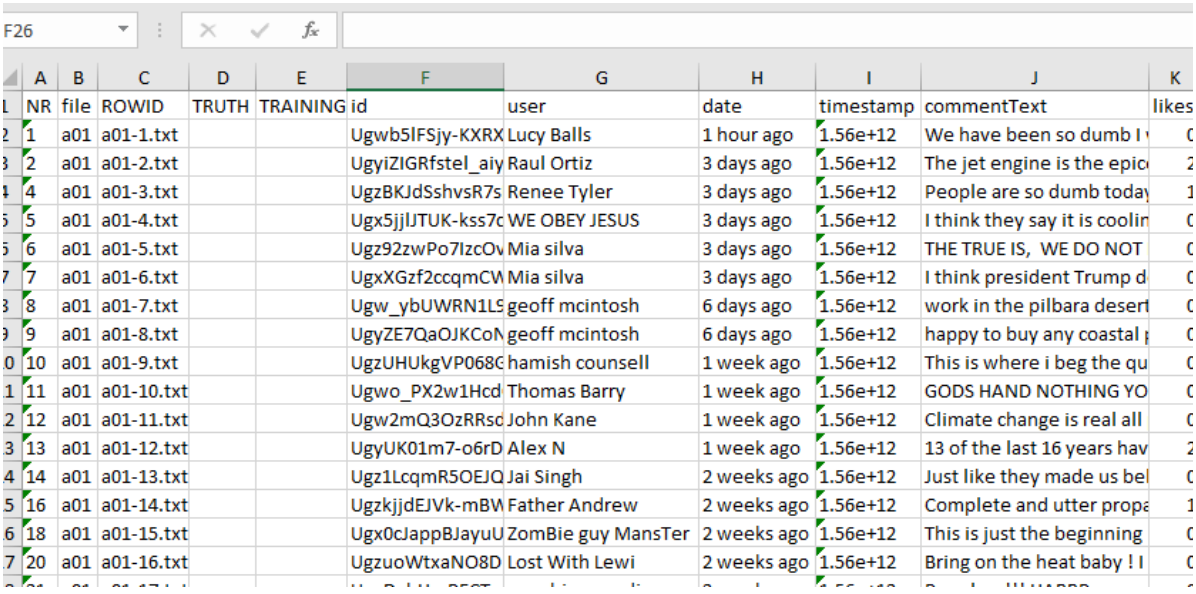
```
#Export the opinions to an excel file
write.xlsx(opinions, "~/ReadMe/climate_change/xlsx/00.xlsx")
```

Script 3: Exporting the file after manual categorization

I continue adding the columns ROWID, TRUTH and TRAINING to the 00.xlsx file. In the column ROWID I have generated a unique identificatory concatenating the number of the row and the value of the column file. With this file I process to capture the categories required.

The 00.xlsx file looks like this:

Representation of the 00.xlsx file



	A	B	C	D	E	F	G	H	I	J	K
L	NR	file	ROWID	TRUTH	TRAINING	id	user	date	timestamp	commentText	likes
2	1	a01	a01-1.txt			Ugwb5lFSjy-KXRX	Lucy Balls	1 hour ago	1.56e+12	We have been so dumb I	0
3	2	a01	a01-2.txt			UgyiZIGRfstel_aiy	Raul Ortiz	3 days ago	1.56e+12	The jet engine is the epic	2
4	4	a01	a01-3.txt			UgzBKJdSshvsR7s	Renee Tyler	3 days ago	1.56e+12	People are so dumb today	1
5	5	a01	a01-4.txt			Ugx5jjIJTUK-kss7c	WE OBEY JESUS	3 days ago	1.56e+12	I think they say it is coolin	0
6	6	a01	a01-5.txt			Ugz92zwPo7IzcOv	Mia silva	3 days ago	1.56e+12	THE TRUE IS, WE DO NOT	0
7	7	a01	a01-6.txt			UgxXGzf2ccqmCW	Mia silva	3 days ago	1.56e+12	I think president Trump d	0
8	8	a01	a01-7.txt			Ugw_ybUWRN1L9	geoff mcintosh	6 days ago	1.56e+12	work in the pilbara desert	0
9	9	a01	a01-8.txt			UgyZE7QaOJKCoN	geoff mcintosh	6 days ago	1.56e+12	happy to buy any coastal p	0
10	10	a01	a01-9.txt			UgzUHUkgVP068C	hamish counsell	1 week ago	1.56e+12	This is where i beg the qu	0
11	11	a01	a01-10.txt			Ugwo_PX2w1Hcd	Thomas Barry	1 week ago	1.56e+12	GODS HAND NOTHING YO	0
12	12	a01	a01-11.txt			Ugw2mQ3OzRRsc	John Kane	1 week ago	1.56e+12	Climate change is real all	0
13	13	a01	a01-12.txt			UgyUK01m7-o6rD	Alex N	1 week ago	1.56e+12	13 of the last 16 years hav	2
14	14	a01	a01-13.txt			Ugz1LcqmR5OEJQ	Jai Singh	2 weeks ago	1.56e+12	Just like they made us bel	0
16	16	a01	a01-14.txt			UgzkjjdEJVk-mBW	Father Andrew	2 weeks ago	1.56e+12	Complete and utter propa	1
18	18	a01	a01-15.txt			Ugx0cJappBJayuU	ZomBie guy MansTer	2 weeks ago	1.56e+12	This is just the beginning	0
20	20	a01	a01-16.txt			UgzuoWtxaNO8D	Lost With Lewi	2 weeks ago	1.56e+12	Bring on the heat baby ! I	0

Figure 8: The file 00.xlsx is used to hand code the categories

Then I proceed to code the categories from the next list:

-1: A negative attitude regarding the content
0: A neutral attitude.
1: A positive attitude regarding content
3: Not relevant for our study
4: Not understandable

Creating the control file

In the next step the file 00.xslx is exported to a dataset, from where it is possible to generate multiple control files that can be useful for different experiments.

With the following script the control file is generated:

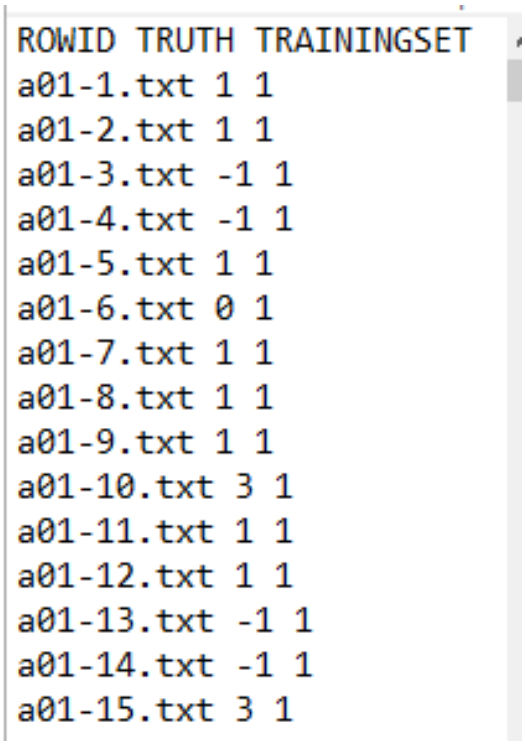
```
install.packages("readxl")
library("readxl")

#import after labeling
opinions_00 <- read.xlsx("~/ReadMe/climate_change/xlsx/00.xlsx", 1)
control_00 <- opinions_00[c(3,4,5)]
write.table(control_00,
file = "~/ReadMe/climate_change/corpus/00/control_00.txt",
sep = " ", row.names = FALSE, quote = FALSE)
```

Script 4: Creating the control file

As a result, I have generated the file “control_00.txt” in my corpus,

The control file created



```
ROWID TRUTH TRAININGSET
a01-1.txt 1 1
a01-2.txt 1 1
a01-3.txt -1 1
a01-4.txt -1 1
a01-5.txt 1 1
a01-6.txt 0 1
a01-7.txt 1 1
a01-8.txt 1 1
a01-9.txt 1 1
a01-10.txt 3 1
a01-11.txt 1 1
a01-12.txt 1 1
a01-13.txt -1 1
a01-14.txt -1 1
a01-15.txt 3 1
```

Figure 9: Representation of the control file created

5.4 Filtering and text pre-processing

To reduce complexity it is optional to use technology to detect no relevant data. Filtering the data frame is optional if applying the HK Method, it helps to reduce complexity, it is also optional to use any kind of technologies or tools for text preprocessing. The method can work without filtering, offering an advantage when The Readme method is implemented. "Our method works without filtering (and in foreign languages), but filters help focus the limited time of human coders on the categories of interest." (Hopkins and King, 2010, p5).

At this point, there are no restrictions to the usage of other methods or software different from the HK literature. Normally, in these kinds of studies, the information excluded is Non-English comments, spam or other kind of comments that do not provide enough information to detect sentiment about the climate change topic. I will apply the HK method applying filters and other standards procedures that reduces complexity. I will experiment with different kinds of preprocessed texts including the opinions with no filter or text procedures.

For the first experiment, I will keep the text with no big changes, later in the next experiment, I procced to filter and pre-process the text following some standards procedures (Welbers, Van Atteveldt and Benoit, 2017).
As mentioned in the methodology, one of the tasks is to make some experiment with datasets with different kinds of pre-processed corpuses.

Experiment 1: Text no preprocessed

Experiment 2: Text cleaned (Puntation, digits and simbols removed)

Experiment 3: stopwords

Experiment 4: Stopwords + text cleaning

Experiment 5: Threefold [.01 : .2]

Experiment 6: Non exhaustive categories present

5.5 Method implementation

After the first steps are completed the training set has been hand coded. The next steps are to read the control file and preprocessing the text before running the readme function to calculate the category proportions. Finally, to calculate the proportions of documents classified into the categories chosen, it is necessary to implements the method running the software ReadMe. This procedure is presented in the next script:

```
setwd(system.file("climate_change/corpus/00", package="ReadMe"))
undergrad.results_00 <- undergrad(control ="control_00.txt", threshold=0.01, python3=F, pyexe=NULL,
sep=" ",printit=FALSE, fullfreq=FALSE)
trainingset_00 <- undergrad.results_00$trainingset[1:100,]
testset_00 <- undergrad.results_00$testset[101:23585,]
undergrad.preprocess <- preprocess(undergrad.results_00)
readme_00 <- readme(preprocess.results_00,  n.subset=300, trainingset= trainingset_00, testset =
testset_00, prob.wt=1, boot.se = TRUE,nboot = 100, printit = FALSE, features = 30)
```

Script 5: The script generalized that implements the HK method

The undergrad function processes the documents according to the control file and stores the data in undergrad.results. When the control.txt file is processed, the data sets and the training sets are stored. The argument “sep” specify the comma-separated argument for the function undergrad.

The function “preprocess” takes the inputted data matrix from undergrad(), removes the columns with variance 0 and store the value in undergrad.preprocess. For windows users the undergrad function throws an error accessing to the file control.txt, this issue is solved in (Tang, 2013). After running this script the function will remove invariant Columns.

The ReadMe function calculates the document category proportion. It also computes bootstrap-based standard errors. The function VA (King and Lu, 2008) that is needed to run the Readme function brings some procedures to the final computation as well.

As a result, the function returns the estimated proportion in each category and other relevant data.

```
#Results t1
readme_00$est.CSMF
readme_00$true.CSMF
readme_00$subsets.est
readme_00$CSMF.se
readme_00$true.CSMF.bootmean
readme_00$true.bootse
readme_00$est.se
readme_00$res.boot
```

Script 6: The code to obtain the results

6. Experiments and analysis

This section contains the experiments realized and the analysis of the results. The first experiment with a known data set is sub-divided into a training set and a test set to perform the implementation of the method and compare the results with known information. The second experiment finds the best performance of the method and in the third experiment I test the method with the entire corpus, estimating the proportions of opinions from activists and deniers.

6.1 Actual data

For the first experiment I have manually classified 600 documents that contains opinions from the YouTube video named "Scientists continue to issue urgent warnings about climate change" (documents from a01-1.txt to a01-600.txt).

Actual data

Category	Value	Opinions	Proportion
Deniers	-1	323	0.537437604
Neutral	0	17	0.02828619
Activists	1	183	0.304492512
No relevant	3	34	0.056572379
No understendable	4	44	0.073211314
Total		**601**	**1**

Table 8: Results from the labeled data on video a01

Percentage of the observed opinions for the video a01

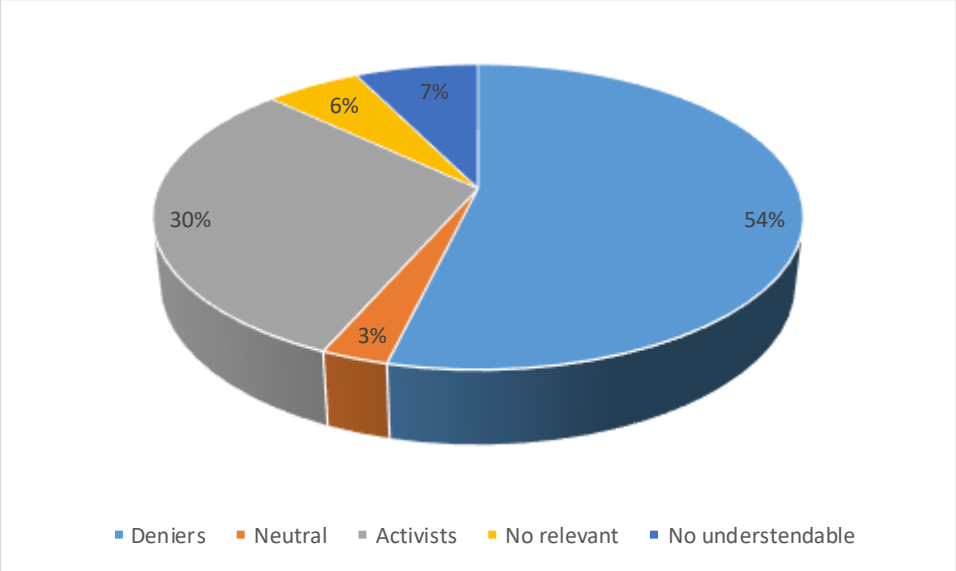


Figure 10: Percentage of the known opinions of the video a01

6.2 Experiment 1

For this first experiment subset a_01 is included that contains the opinions from a video posted by an activist. In the subset a_01 to 600 the opinions has been delimited as all of them was classified manually. This experiment will compare the results of the observed data to the results of the estimated proportion with the features described below. The model will be validated through k-folk cross validation and bootstrapping.

For the experimentation phase I am going to use only the opinion from the video number 1 from activists. I proceed to create a control file for this propose.

```
opinions_a01 <- opinions_00[opinions_00$file == "a01",]
control_a01 <- opinions_a01[c(3,4,5)]
write.table(control_a01,
file = "~/ReadMe/climate_change/corpus/00/control_a01.txt",
sep = " ", row.names = FALSE, quote = FALSE)
```

Script 7: Script to generate the control file to the experimentation phase

List of features for the experiment 1

Characteristics of the dataset

Characteristics	**Value**
Corpus name (subset)	A_01
Control file	Control_a01_1.txt
Corpus size	600 documents
Training set	100
Test set	500
Random training set	No

Preprocessing features

Features	**Value**
Threshold	0.01
N-Grams	Unigrams
Full Frequency	No
Case Sensitive	Ignored

Word features	30
Probability of weights	1
Stop words	NO
Preprocessing and text cleaning	Standard in HK Method
Other Special features	No

Methods for validation

Validation	**Type**
Validation	K fold cross validation K=10
Standard error	Bootstraping n=100

Implementation for the first experiment

```
setwd(system.file("climate_change/corpus/00", package = "ReadMe"))
undergrad_a01 <- undergrad(control = "control_a01_1.txt", stem = T, threshold =.01 , printit=FALSE,
fullfreq = FALSE)
preprocess.undergrad_a01_1 <- preprocess(undergrad_a01_1)
readme_a01_k1 <- readme(preprocess.undergrad_a01_1,  n.subset=600, prob.wt=1, boot.se =
FALSE,nboot = 100, printit = FALSE)
```

Script 8: Script that implements the experiment 1

Cross validation implementation

This code implements the HK method for K-fold K=1, The same code is repeated for the next K fold, the control file must be changed with the new values for the training set and test set

```
#implements the K Fold k=1
setwd(system.file("climate_change/corpus/00", package = "ReadMe"))
undergrad_a01_k1 <- undergrad(control = "control_a01_k1.txt", stem = T, threshold =.01 ,
printit=FALSE, fullfreq = FALSE)
preprocess.undergrad_a01_k1 <- preprocess(undergrad_a01_k1)
readme_a01_k1 <- readme(preprocess.undergrad_a01_k1,  n.subset=600, prob.wt=1, boot.se =
FALSE,nboot = 100, printit = FALSE, features = 30)

#Results
readme_a01_k1$est.CSMF
readme_a01_k1$subsets.est
```

Script 8: Script that implements the k fold Cross Validation

6.3 Analysis 1

The results from the nonparametric method shows a relatively good accuracy if we compare with the real data, on this first instance, the HK method does not present important biases, but it does not present favorable results as other studies have shown (D. J. Hopkins and King, 2010).

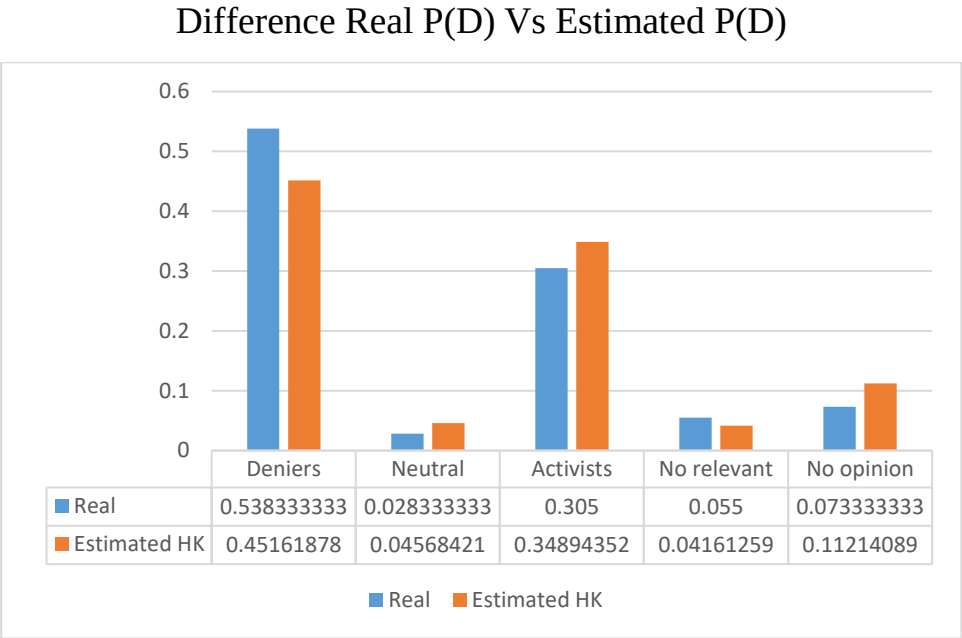


Figure 11: Shows the difference in proportions from the estimated data compared with the real data

On the figure 11, we can observe low biases in the proportions estimated. It is a nice result for a confidence interval of 95%, we can denote that the opinions from deniers has been estimated as opinions for activists. It is a good point to consider in a detailed analysis of the corpus.

The next graphic shows the distribution of the estimated values paired with the real data.

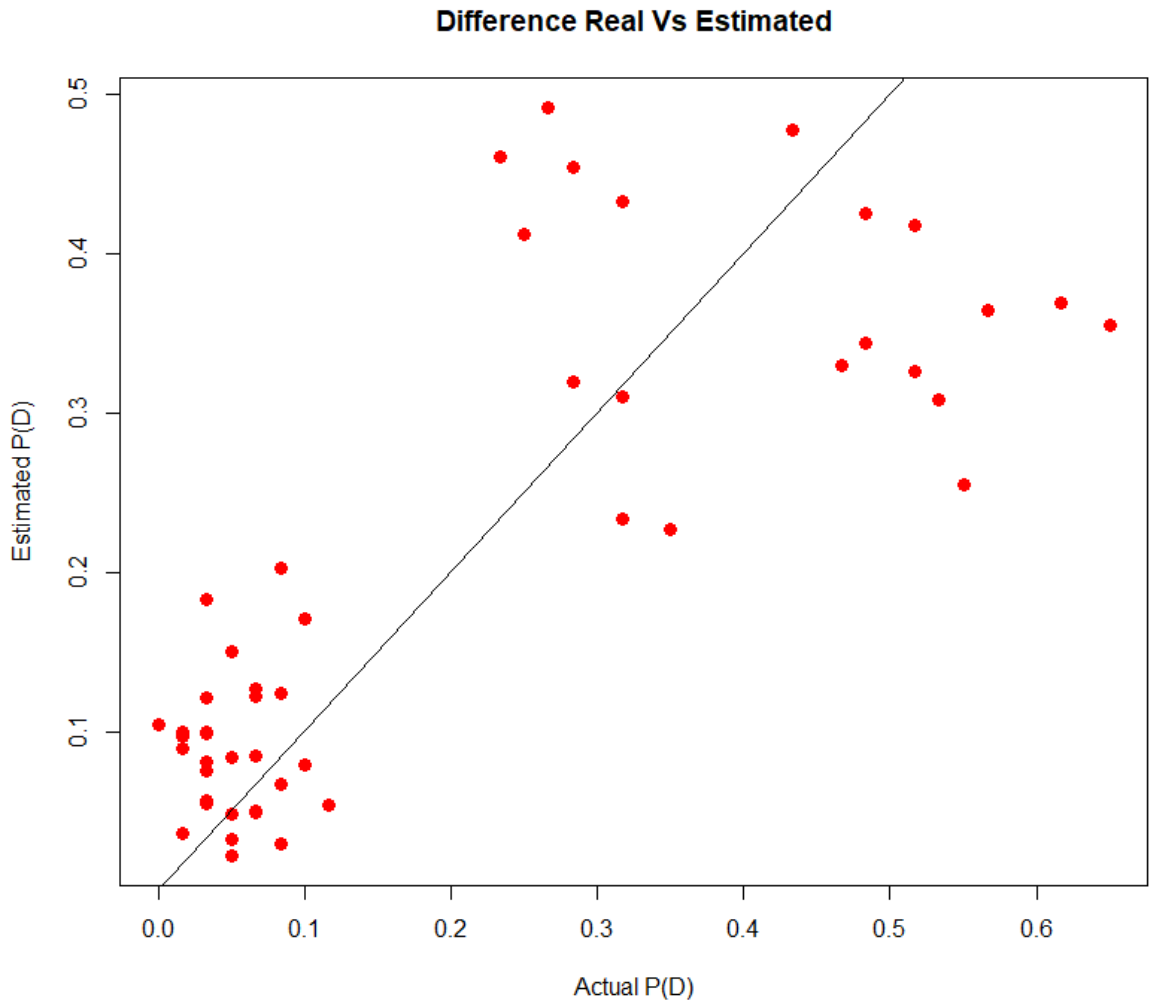


Figure 12 shows the distribution of the estimated proportion when are compared with known proportion

Bootstrap Standard error

The bootstrap technique calculates the standard error between the known data and the estimated data. In the experiment one the number of iterations of the method was 100.

The graphic of the calculated proportions

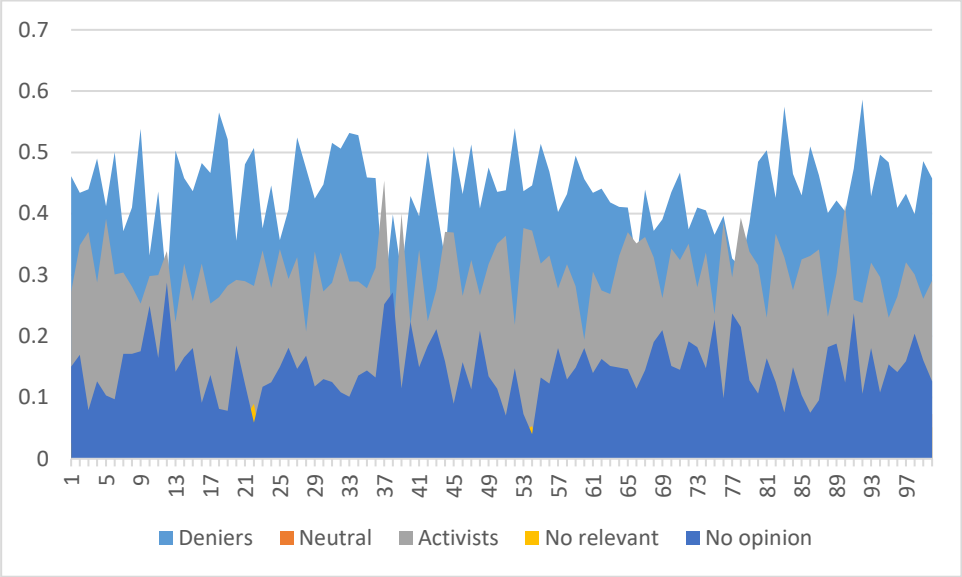


Figure 13: The graphic shows the total of the 100 proportions estimated.

As showed in the figure 12, the difference between proportions are wide, the method present presents important issues for the first experiment. Is recommendable to make adjusts in some features as filtering or text preprocessing.

Most relevant values obtained from bootstrap

Iteration	Deniers	Neutral	Activists	No Relevant	No opinion
12	0.2805489	0.043366227	0.3391365	0.049499618	0.2874487
37	0.2679015	0.026008229	0.4538582	0	0.25223214
38	0.3986763	0.114380987	0.188955	0.026654966	0.27133274
92	0.5862206	0.031149212	0.2541419	0.022342075	0.10614621
Real	**0.538333333**	**0.028333333**	**0.305**	**0.055**	**0.073333333**

Table 9: The table shows the most relevant bias presented from the 100 interactions.

The table shows the most relevant values from the bootstrap iteration. We can observe in the iteration 92, the maximum proportion for deniers', It no represent biases compared with the real values. We can denote that the minimum value from activists' opinions come in the interaction number 38, there the biases goes to the "No opinions" category. The same happens in the iteration number 37, there is found the minimum for deniers, the maximum for activists

and an important tendency to “No Opinions”. Finally, we can observe the maximum for “No opinions” in the interaction 12, carrying an important bias for the “Deniers” proportion.

Validating the HK method according with the K Folk approach. The number of interactions was 10 with a training set of 540 known opinions to be tested in a dataset of 60 opinions. The K Folk cross validation shows also important bias for the method, we can observe on the table below an important bias in K = 4. The values for Deniers are very low, while the No Opinions category get most of the biased proportion. In all the interactions the value for deniers are low compared with the real data, for the activist´s category the proportions trends to be higher than the real proportions. On this point we know that the method needs some adjusts.

10 Fold Cross Validation for the experiment 1

k	Deniers	Neutral	Activists	NR	NO
1	0.36933172	0.10440927	0.41158097	0.04773171	0.06694633
2	0.41733848	0.07521134	0.23326699	0.15016845	0.12401474
3	0.32971958	0.08917324	0.47764361	0.02240806	0.0810555
4	0.25448661	0.09719171	0.31910602	0.12686591	0.20234976
5	0.34333881	0.0846879	0.30967922	0.07924228	0.18305179
6	0.42547855	0.05662357	0.22688433	0.12072051	0.17029304
7	0.32575652	0.10005613	0.49119747	0.02952028	0.0534696
8	0.30821612	0.05432191	0.43227315	0.08330321	0.12188561
9	0.36407575	0.09988107	0.45384305	0.0319686	0.05023153
10	0.3545879	0.09901619	0.46049643	0.03645138	0.04944809
Real	**0.538333333**	**0.02833333**	**0.305**	**0.055**	**0.07333333**

Table 10: Results in proportions obtained from the k-fold cross validation

6.4 Experiment 2

On the last experiment we found that the HK method does not performs very well when it is working with opinions from climate change, for the experiment 2, I will build a new corpus where I will be using some preprocess features and adding the username.

I will experiment with different values for threshold [0.00 to 0.05], to try to find the best performance for the method. For the feature threshold, when setting values over 0.05, the function readme throws an error because the sample is too low. I will also add the features of "full frequency" and amplifying the word features to 50.

In the experiment 1 there was not any correlation between the size of the training set. We have to remember that for the experiment we used a training set of 100 observed opinions and for the validation we used a training set of 540 documents. That difference did not make a clear difference in the performance. For the experiment 2 it will be not necessary to experiment with different sizes for the training set.
During the labeling phase I found that in some situations when it was difficult to determine the type of opinion, it was easier to understand the opinion when I read other opinions from the same user. The same kind of hint could help the machine to predict the category of the document.

I convert the username "example user" to "@example_user", adding the "@" symbol before the name and replacing space for "_". The new username was concatenated to the new text.

Examples for the new format of comments:

File	Comment
a01-1.txt	@Lucy_Balls We have been so dumb I wish we could start over
a01-4.txt	@WE_OBEY_JESUS I think they say it is cooling now Its not science
a01-16.txt	@Lost_With_Lewi Bring on the heat baby I love hot summers
a01-17.txt	@noushin_saeedi People HARRP Look it up

Then a new corpus was built with the name corpus_02, from this corpus I will use a subset with only the opinions from the video a01

List of characteristics in experiment 2

Characteristics of the dataset

Characteristics	**Value**
Corpus name (subset)	Corpus_02_a01
Control file	Control_a01_1.txt
Corpus size	600 documents
Training set	100
Test set	500
Random training set	No

Preprocesing features

Features	**Value**
Threshold	[0.00 : 0.05]
N-Grams	Unigrams
Full Frequency	Yes
Case Sensitive	Ignored
Word features	50
Probability of weights	1
Stop words	Yes
Preprocessing and text cleaning	Quanteda R package
Other Special features	Username added to the documents

Validation methods

Validation	**Type**
Validation	K fold cross validation K=10
Standard error	Bootstraping no available

6.5 Analysis 2

The second analysis involves the implementation of the experiment 2, running the method with new features and a threshold variable from .00 to .05. On this experiment it was not possible to implement the bootstrapping technique to find the standard error. This operation demanded high hardware resources that were not available.

After the experiment with multiple variables in threshold, it was possible to obtain the best performance for the method for this kind of corpus.

Estimated proportions when the threshold is variable

Test	Threshold	Deniers	Neutral	Activists	NR	NO
1	0.00	0.309937412	0.062852482	0.383770034	0.127394128	0.116045945
2	0.01	0.467834679	0.015023694	0.334811028	0.027902353	0.154428244
3	0.02	0.614578018	0.081557061	0.241709737	0.009549089	0.052606096
4	0.03	0.503426115	0.043422816	0.222601522	0.175615311	0.054934238
5	0.04	0.454383224	0.027162764	0.342018183	0.061023433	0.115412396
Real		**0.538333333**	**0.028333333**	**0.305**	**0.055**	**0.073333333**

Table 11: The table n shows the mean of the estimated proportions when the readme function was implemented 10 times. In every test the method doesn't present variances between the interactions.

The method has got better results after the new features are implemented, the lowest difference between proportions is found where the values of threshold is 0.04. This is a good result that must be validated through 10 fold cross validation.

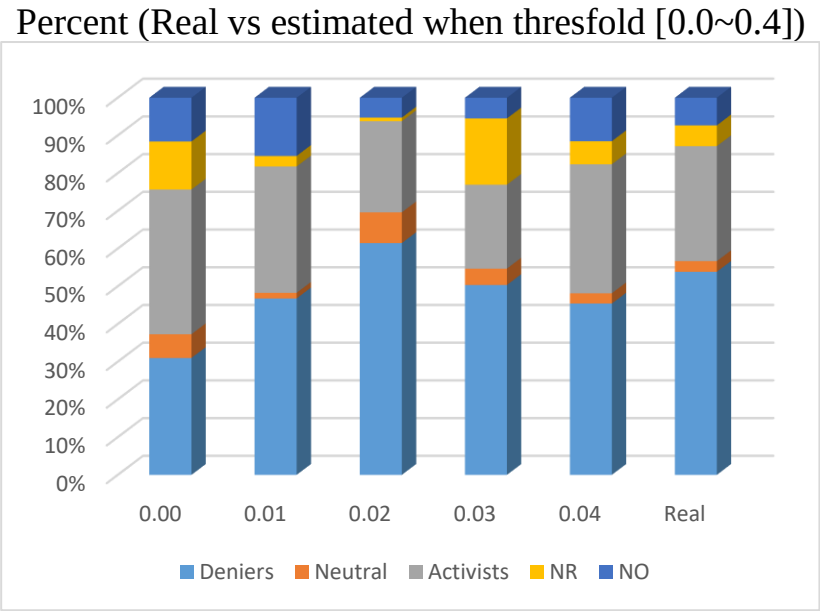


Figure 14: The method shows similitudes in threshold [0.04] when is compared with re known data.

The figure 13, shows good accuracy of the HK method between the Threshold in 0.04 and the real data. This result was tested in 10 iterations with very low standards error. These results are showed in the table below:

Proportions for the test with threshold = 0.04

	Deniers	**Neutral**	**Activists**	**NR**	**NO**
1	0.44491536	0.026241	0.358113	0.045981	0.124749
2	0.45312423	0.023997	0.354144	0.052058	0.116676
3	0.45160187	0.026502	0.353228	0.046384	0.122284
4	0.45376072	0.027777	0.356162	0.048056	0.114244
5	0.44967166	0.026411	0.35121	0.048123	0.124585
6	0.43773052	0.026755	0.361608	0.049478	0.124428
7	0.43311106	0.021958	0.371863	0.04435	0.128719
8	0.50254815	0.042551	0.222291	0.179662	0.052948
9	0.46542681	0.023102	0.337387	0.049629	0.124455
10	0.45194186	0.026333	0.354177	0.046511	0.121036

Table 12: Estimation in proportion for threshold = 0.04

K-Fold Cross validation

It was not possible to implement the cross-validation K fold for K=10; The software did not run the function for some folds due to the sample being too small. The cross validation was implemented with fold K=5, with no warnings:

Difference in proportions (Mean)

Threshold	Deniers	Neutral	Activists	NR	NO	Dif
0.00	0.4242649	1.2183229	0.2582624	1.3162569	0.5824447	3.7995517
0.01	0.1309573	0.469752	0.0977411	0.4926845	1.1058397	2.2969745
0.02	0.141631	1.8784845	0.2075091	0.8263802	0.2826441	3.3366489
0.03	0.0648431	0.53257	0.2701589	2.1930057	0.2508968	3.3114745
0.04	**0.1559445**	**0.0413142**	**0.1213711**	**0.109517**	**0.5738054**	**1.0019521**

Table 13: Difference real vs estimated in k5 cross validation

Graphic of the difference real vs estimated in K5-fold cross-validation

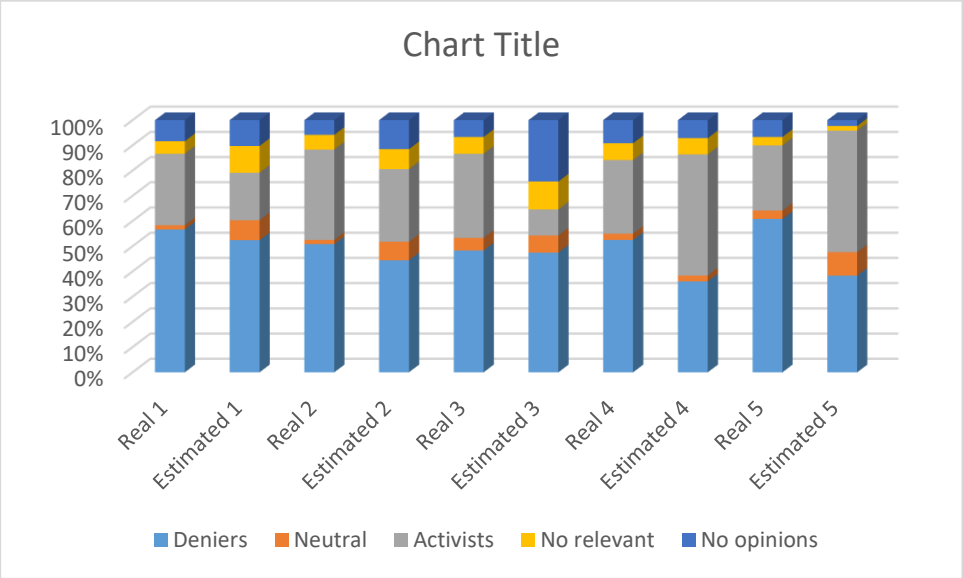


Figure 15: Comparation of experiment 2 vs real (K5 Cross validation)

As showed in the previous results, the method still present biases as showed in the figure 14. If we look at the difference in the test number 4, we denote that the method does not performs well when estimating small subset. The same occurs with the test number 3. If the subset is relatively small the software shows some warning messages as “the sample could be too low”.

Comparative chart between real data and results from 5 fold validation

real	Deniers	Neutral	Activists	No relevant	No opinions
Real 1	0.566666667	0.016666667	0.283333333	0.05	0.083333333
Estimated 1	0.52395733	0.07925925	0.18834773	0.10604534	0.10239035
Real 2	0.508333333	0.016666667	0.358333333	0.058333333	0.058333333
Estimated 2	0.44429369	0.0734919	0.28802217	0.07911592	0.11507631
Real 3	0.483333333	0.05	0.333333333	0.066666667	0.066666667
Estimated 3	0.47471527	0.06793568	0.10311762	0.11077764	0.24345379
Real 4	0.525	0.025	0.291666667	0.066666667	0.091666667
Estimated 4	0.36035022	0.02373765	0.48011479	0.06453471	0.07126263
Real 5	0.608333333	0.033333333	0.258333333	0.033333333	0.066666667
Estimated 5	0.38359441	0.0928118	0.4818238	0.01867554	0.02309444

Table 14: Comparation table for the actual data vs the estimated when threshold is 0.04

For now, the best performance of the HK method occurs when some features are implemented, including setting up the feature threshold on 0.04.

Implementing the experiment 2, involved more quantity of data to be analyzed, the number of words (features in the readme function) to analyze was changed from 30 to 50. It caused computing time consuming and throws constantly error due to low RAM memory.

7. Conclusion

This thesis implemented the nonparametric method of Hopkins and King to measure opinions on the social media YouTube, where the goal was to explore this method when used to measure opinions about climate change.

After analyzing the results, it can be concluded that the method of Hopkins and King is not giving good accuracy when the subset is small. When the subset had the size of 500 with a training set of 100, the estimation of the proportions was more accurate. When the features for the function ReadMe increased, the ReadMe software brings better results, but this implies the use of more resources (time and hardware). Finally, it was impossible to implement the method in all the data (more than 83,000 documents) when the function readme was implemented.

It is possible to obtain better results in future researches adapting changes in the corpus, experiment with different tasks for text preprocessing or filtering, as example dropping all the opinions that contains difficulties to understand by the human, adding some special tags to the text with clear tendency, or another techniques that are useful according with the field of text mining, the hardest part is to identify automatically where to apply the filters or where to add extra tags on this kind of information.

Unfortunately, the kind of data used to estimate the type of opinion was complicate to manage manually and then it is even more complicated to analyze for the computer. In this study it was clear that the classification of opinions is a hard task for humans as well as for the most sophisticated methods of text analysis. Tagging performed by humans is expensive in terms of time and energy but nevertheless indispensable. Humans are still better at recognizing sarcasm, colloquialisms or the continuity of certain conversations for example when a person reappears in the debate after having been absent for some time.

My conclusion is that the comments on YouTube or other social media platforms represented a difficult classification task for coders. The method of HK showed bad performance in assisting the human effort when few features was selected, the method show better results after some adjustments, but that represented long computing time to analyze opinions.

References

van ’t Haar, G. *et al.* (2013) *Between Broadcasting Political Messages and Interacting With Voters, Information, Communication & Society*.

Alpaydın, E. (2014) ‘Introduction to machine learning’, *Methods in Molecular Biology*, 1107, pp. 105–128. doi: 10.1007/978-1-62703-748-8-7.

Cavnar, W. B., Trenkle, J. M. and Mi, A. A. (1994) ‘N-Gram-Based Text Categorization’, *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp. 161–175. doi: 10.1.1.53.9367.

Cazau, P. (2006) ‘Fundamentos de Estadística’. Available at: http://greco.uclm.es/Record/Xebook1-2048.

Ceron, A. *et al.* (2014) ‘Every tweet counts ? How sentiment analysis of social media can improve our knowledge of citizens ’ political preferences with an application to Italy and France’. doi: 10.1177/1461444813480466.

Ceron, A., Curini, L. and Iacus, S. M. (2014) ‘Using social media to forecast electoral results. A meta-analysis’, *UNIMI-Research Papers in …*, 25(3).

Faizi, R., El Afia, A. and Chiheb, R. (2013) ‘Exploring the Potential Benefits of Using Social Media in Education’, *International Journal of Engineering Pedagogy (iJEP)*, 3(4), p. 50. doi: 10.3991/ijep.v3i4.2836.

Frank, E., Hall, M. A. and Witten, I. H. (2016) ‘WEKA Workbench’.

Gooding-williams, R. (2017) ‘Stanford Encyclopedia of Philosophy Stanford Encyclopedia of Philosophy’, (September).

Grimmer, J. and Stewart, B. M. (2013) ‘Text as data: The promise and pitfalls of automatic content analysis methods for political texts’, *Political Analysis*, 21(3), pp. 267–297. doi: 10.1093/pan/mps028.

Hambrick, M. E. *et al.* (2010) ‘Understanding Professional Athletes’ Use of Twitter: A Content Analysis of Athlete Tweets’, *International Journal of Sport Communication*, 3(4), pp. 454–471. doi: 10.1123/ijsc.3.4.454.

Henri Cohen and claire lefebvre (2005) *Handbook of categorization in cognitive science*.

Hopkins, D. *et al.* (2012) ‘ReadMe: Software for automated content analysis’, (617), p. 10. Available at: http://faculty.washington.edu/jwilker/559/Readme documentation.pdf.

Hopkins, D. J. and King, G. (2010) ‘A method of automated nonparametric content analysis for social science’, *American Journal of Political Science*, 54(1), pp. 229–247. doi: 10.1111/j.1540-5907.2009.00428.x.

Hopkins, D. and King, G. (2010a) ‘A Method of Automated Nonparametric Content Analysis for Social Science’, *American Journal of Political Science*, 54(1), pp. 229–247. doi: 10.1111/j.1540-5907.2009.00428.x.
Hopkins, D. and King, G. (2010b) ‘Extracting systematic social science meaning from text’, *American Journal of Political Science*, 54(1), pp. 229–47. Available at: http://polmeth.wustl.edu/retrieve.php?id=701%5Cnhttp://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.179.29&rep=rep1&type=pdf.
J. Clement (2019) *Twitter: number of monthly active users 2010-2019*. Available at: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/). (Accessed: 5 May 2019).
King, G. (2007) ‘Extracting Systematic Social Science Meaning from Text ({&} Cause-Specific Mortality Rates from Symptom Data) References Daniel Hopkins and Gary King . “ Extracting Systematic Social Science’.
King, G. (2016) ‘An Improved Method of Automated Nonparametric Content Analysis for Social Science 1 Mortality Data , Developed Countries ’:, (617).
King, G. and Lu, Y. (2008) ‘Verbal Autopsy Methods with Multiple Causes of Death’, 23(1), pp. 78–91. doi: 10.1214/07-STS247.
Klostermann, P. (2015) *YouTube Comment Scraper*. Available at: https://github.com/philbot9/youtube-comment-scraper (Accessed: 6 May 2019).
Krippendorff, K. (2003) ‘Content Analysis: An Introduction to Its Methodology’, *Content Analysis: An Introduction to Its Methodology*, p. 440. doi: 10.2307/2288384.
Martinez, J. (2019) *Climate Change YouTube comments*. Available at: https://github.com/PepeBergen09/Climate_Change_YouTube_Comments.
Melendez, S. *et al.* (2018) *ReadMe: Software for Automated Content Analysis, 2010*. Available at: https://gking.harvard.edu/readme.
Nunez-Mir, G. C. *et al.* (2016) ‘Automated content analysis: addressing the big literature challenge in ecology and evolution’, *Methods in Ecology and Evolution*, 7(11), pp. 1262–1272. doi: 10.1111/2041-210X.12602.
Pang, B. and Lee, L. (2008) ‘Opinion mining and sentiment analysis’, *Foundations and Trends in Information Re- trieval*, 2(1).
Pang, B., Lee, L. and Vaithyanathan, S. (2002) ‘Thumbs up ? Sentiment Classification using Machine Learning Techniques’, *Proceedings of EMNLP 2002*, (July), pp. 79–86.
Prasad, B. D. (2008) ‘Content Analysis A method in Social Science Research’, *Research Methods for Social Work*, (2008), pp. 173–193. Available at:

http://repository.upenn.edu/cgi/viewcontent.cgi?article=1232&context=asc_papers&sei-redir=1&referer=http://scholar.google.com.my/scholar_url?hl=en&q=http://repository.upenn.edu/cgi/viewcontent.cgi?article=1232&context=asc_papers&sa=X&scisig=AAGBfm23DWDQYw.

Ronen Feldman; James Sanger; (2007) *The Text Mining Handbook*. doi: 10.1017/CBO9780511546914.

Samuel, A. L. (1959) 'Some Studies in Machine Learning Using the Game of Checkers', *IBM Journal of Research and Development*, 3(3), pp. 210–229. doi: 10.1147/rd.33.0210.

Singh, V. and Dubey, S. K. (2014) 'Opinion Mining and Analysis: A Literature Review', *2014 5Th International Conference Confluence the Next Generation Information Technology Summit (Confluence)*, pp. 232–239. doi: 10.1109/CONFLUENCE.2014.6949318.

Soleman, N., Chandramohan, D. and Shibuya, K. (2006) 'Verbal autopsy: Current practices and challenges', *Bulletin of the World Health Organization*, 84(3), pp. 239–245. doi: 10.2471/BLT.05.027003.

Sterne, J. (2010) *Social Media Metrics: How to Measure and Optimize Your Marketing Investment*.

Talib, R. *et al.* (2016) 'Text Mining: Techniques, Applications and Issues', *IJACSA) International Journal of Advanced Computer Science and Applications*, 7(11), pp. 414–418. Available at: www.ijacsa.thesai.org.

Tang, C. (2013) *[readme] Python error on demo(clinton) (Windows Fix)*. Available at: https://lists.gking.harvard.edu/pipermail/readme/2013-October/000028.html (Accessed: 5 October 2019).

The R Project for Statistical Computing (2018). Available at: https://www.r-project.org.

UMSL, U. C. S. (2004) 'An Introduction to Content Analysis', *Colorado State University*, p. 2. doi: 10.2307/2288384.

Welbers, K., Van Atteveldt, W. and Benoit, K. (2017) 'Text Analysis in R', *Communication Methods and Measures*, 11(4), pp. 245–265. doi: 10.1080/19312458.2017.1387238.

World Health Organization (WHO) (2012) 'Verbal autopsy standards', pp. 1–143.

YouTube (2019) *No Title*. Available at: http://www.youtube.com.

Appendix

A. List of Software and tools Used

R-Cran project: R is a free software that provides a wide variety of statistical and graphical techniques. R as a programming language is widely used for research in statistical methodology.

RStudio is an IDE (integrated development environment) for R, It is a friendly alternative where can be used console, source, plots, workspace, help, history, etc

R - VA Package: VA is an easy-to-use R program that automates the analysis of verbal autopsy data. Some VA functions performs tasks of the ReadMe Package

The ReadMe software computes a set of text documents into multiple categories chosen by the user, is needed the hand code classification. ReadMe will report the proportion of documents within each of the given categories.

Textcat is an R Extension package for n-gram based text categorization that implements the Cavnar and Trenkle approach.

R Package Weka / Weka 3.8 Windows version: Weka is a collection of machine learning algorithms and tools for data processing that can try out existing methods offering flexibility. (Frank, Hall and Witten, 2016) This Package contains more than 100 algorithms for classification, 75 for data processing. This is a good friendly alternative to R.

Octoparse

YouTube comments Scraper

B. Software readme Requirements and installation

Requiriments

- Requires Python and R
- Operative system Linux, Mac or Windows

Installation script on Linux

```
> mkdir ~/.R ~/.R/library
> R_LIBS = "~/.R/library"
> install.packages("VA", repos= "http://r.iq.harvard.edu", type="source")
> install.packages("ReadMe", repos = "http://r.iq.harvard.edu", type="source")
```

C. List of videos analyzed

C.1 Videos from activists

1. Title: Scientists continue to issue urgent warnings about climate change | 7.30
   URL: https://www.youtube.com/watch?v=Bc8sppzaueo
   Author: ABC News

2. Title: Greta Thunberg's emotional speech to EU leaders
   URL: https://www.youtube.com/watch?v=FWsM9-_zrKo
   Author: Guardian News

3. Title: How We Can Make the World a Better Place by 2030 | Michael Green | TED Talks
   URL: https://www.youtube.com/watch?v=o08ykAqLOxk
   Author: TED

4. Title: Fleeing climate change - the real environmental disaster | DW Documentary
   URL: https://www.youtube.com/watch?v=cl4Uv9_7KJE
   Author: DW Documentary

5. Title: Climate Change - The Facts
   URL: https://www.youtube.com/watch?v=0ypaUH57MO4
   Author: Explore512

6. Title: Climate Change: "If we lose the Arctic, we lose the whole world" (w/ Guy McPherson)
   URL: https://www.youtube.com/watch?v=3SPiXBSjc-4
   Author: Thom Hartmann Program

7. Title: Causes and Effects of Climate Change
   URL: https://www.youtube.com/watch?v=G4H1N_yXBiA
   Author: National Geographic

8. Title: Why we're heading for a 'climate catastrophe'
   URL: https://www.youtube.com/watch?v=pJ1HRGA8g10&t=4s
   Author: BBC Newsnight

9. Title: ESA and climate change
   URL: https://www.youtube.com/watch?v=ezAZ5WVAOyI
   Author: European Space Agency, ESA

10. Title: The battle against climate change by Paul Kingsnorth
    URL: https://www.youtube.com/watch?v=Q_s8Vo00Xug
    Author: vpro documentary

11. Title: The Real National Emergency Is Climate Change: A Closer Look
    URL: https://www.youtube.com/watch?v=mC4bYqbQihI
    Author: Late Night with Seth Meyers

12. Title: DeGrasse Tyson: We have to believe science on climate change
    URL: https://www.youtube.com/watch?v=Jm_YoL9ykC4
    Author: CNN

13. Title: Paris Agreement: Last Week Tonight with John Oliver (HBO)
    URL: https://www.youtube.com/watch?v=5scez5dqtAc
    Author: Last Week Tonight with John Oliver (HBO)

14. Title: Heart-Wrenching Video: Starving Polar Bear on Iceless Land
    URL: https://www.youtube.com/watch?v=_JhaVNJb3ag
    Author: National Geographic

15. Title: An Emotional, Powerful Speech On Climate Change
    URL: https://www.youtube.com/watch?v=7SSXLIZkM3E
    Author: The Daily Conversation
16. Title: Climate Change: The State of the Science
    URL: https://www.youtube.com/watch?v=_EWOrZQ3L-c
    Author: International Geosphere-Biosphere Programme

17. Title: How climate change makes hurricanes worse
    URL: https://www.youtube.com/watch?v=_0TCrGtTEQM
    Author: Vox

18. Title: David Attenborough: 'Climate Change - Britain Under Threat'
    URL: https://www.youtube.com/watch?v=Cq1oFhTINXE
    Author: Carbon Control

19. Title: Climate Change: What Happens If The World Warms Up By 2°C?
    URL: https://www.youtube.com/watch?v=9GjrS8QbHmY
    Author: Sky News

20. Title: A simple and smart way to fix climate change | Dan Miller | TEDxOrangeCoast
    URL: https://www.youtube.com/watch?v=0k2-SzlDGko
    Author: TEDx Talks

C.2 Videos from deniers

1. Title: What They Haven't Told You about Climate Change
   URL: https://www.youtube.com/watch?v=RkdbSxyXftc
   Author: PragerU

2. Title: DEBUNKED: Top 5 "Climate Change" Myths
   URL: https://www.youtube.com/watch?v=QwviDPo4Rh4
   Author: StevenCrowder

3. Title: GLOBAL WARMING IS THE BIGGEST FRAUD IN HISTORY - Dan Pena
   URL: https://www.youtube.com/watch?v=NjlC02NsIt0
   Author: London Real

4. Title: Nobel Laureate in Physics; "Global Warming is Pseudoscience"
   URL: https://www.youtube.com/watch?v=SXxHfb66ZgM
   Author: 1000frolly

5. Title: Nobel Laureate Smashes the Global Warming Hoax
   URL: https://www.youtube.com/watch?v=TCy_UOjEir0
   Author: 1000frolly

6. Title: The Biggest Lie About Climate Change
   URL: https://www.youtube.com/watch?v=TbW_1MtC2So&t=46s
   Author: AsapSCIENCE

7. Title: DEBUNKED: Great Lakes Climate Change Hysteria! | Louder With Crowder
   URL: https://www.youtube.com/watch?v=CJBrJRCXJmA
   Author: StevenCrowder

8. Title: WHY I SAID GLOBAL WARMING IS THE BIGGEST FRAUD IN HISTORY - Dan Pena
   URL: https://www.youtube.com/watch?v=m0sY2tjmr_Y
   Author: London Real

9. Title: Climate Change in 12 Minutes - The Skeptic's Case
   URL: https://www.youtube.com/watch?v=0gDErDwXqhc
   Author: Stefan Molyneux

10. Title: Noam Chomsky: How Climate Change Became a 'Liberal Hoax'
    URL: https://www.youtube.com/watch?v=FJUA4cm0Rck
    Author: The Nation

11. Title: Busting Climate Change Myths | Answers With Joe
    URL: https://www.youtube.com/watch?v=sZB1YtQtHjE
    Author: Joe Scott

12. Title: Global Warming Is A Hoax
    URL: https://www.youtube.com/watch?v=-AwNKQqLESc
    Author: Counter Arguments

13. Title: The Global Warming Hoax Explained for Dummies
URL: https://www.youtube.com/watch?v=nq4Bc2WCsdE
Author: MrJacktemplar

14. Title: Lord Christopher Monckton - Global Warming is a Hoax
URL: https://www.youtube.com/watch?v=UGqcweY1a3I
Author: ideacity

15. Title: Former NASA Scientists... Global Warming Hoax
URL: https://www.youtube.com/watch?v=aEaFzhoS67I
Author: PatriotNetworkAZ

16. Title: Global Warming Hoax, Best Document Ever
URL: https://www.youtube.com/watch?v=DJBDI7jVMqM
Author: seawapa.org

17. Title: Global Warming Hoax, Planned in 1961
URL: https://www.youtube.com/watch?v=SvcuylMrkXk
Author: eenkmouse2311

18. Title: The Climate Change Hoax, with Professor Willie Soon at Camp Constitution 7-3-17
URL: https://www.youtube.com/watch?v=4YMttEhtgpk
Author: Camp Constitution

19. Title: Donald Trump Believes Climate Change Is A Hoax | All In | MSNBC
URL: https://www.youtube.com/watch?v=yqgMECkW3Ak
Author: MSNBC

20. Title: The experts explain the global warming myth: John Coleman
URL: https://www.youtube.com/watch?v=AA3OA_2S4QY
Author: KUSI News

D. General stats of the analyzed videos

Activists stats videos

Nr	Published	Analyzed	Duration	Views	Likes	Dislikes	Comments
1	13.12.2018	06.05.2019	6:30	92,063	1200	288	1861
2	16.04.2019	06.05.2019	4:11	341,278	1400	1000	2097
3	03.11.2015	06.05.2019	14:39	576,365	7000	368	591
4	01.05.2019	06.05.2019	42:25	42,105	733	190	555
5	19.04.2019	06.05.2019	57:31	364,560	6200	288	1762
6	03.05.2019	06.05.2019	11:22	19,420	801	37	461
7	28.08.2017	06.05.2019	3:04	804,107	7600	382	951
8	08.10.2018	06.05.2019	15:20	282,082	3800	614	4045
9	20.03.2019	06.05.2019	4:30	31,341	859	157	484
10	26.04.2019	06.05.2019	49:32	35,067	942	149	550
11	20.02.2019	06.05.2019	7:59	1 974,510	22000	660	3710
12	14.10.2018	06.05.2019	9:26	270,448	4900	436	2728
13	04.06.2017	06.05.2019	20:57	1 1379,711	162000	12000	14381
14	11.12.2017	06.05.2019	1:22	2 005,992	22000	3700	8831
15	11.11.2013	06.05.2019	4:06	1 219,108	16000	162	702
16	19.11.2013	06.05.2019	4:04	837,197	2600	190	664
17	28.08.2017	06.05.2019	3:22	675,222	17000	1200	2645
18	07.12.2013	06.05.2019	1:00:14	308,171	1200	172	985
19	29.11.2015	06.05.2019	2:35	220,528	1400	150	373
20	23.10.2014	06.05.2019	16:31	213,306	2700	322	1127

Deniers stats videos

Nr	Published	Analyzed	Duration	Views	Likes	Dislikes	Comments
1	27.07.2019	06.05.2019	4:54	2 645,669	35000	20000	13052
2	30.08.2016	06.05.2019	20:04	1 961,939	57000	14000	23679
3	28.12.2017	06.05.2019	5:44	1 589,376	33000	7900	17787
4	17.12.2015	06.05.2019	31:38	1 501,141	23000	3100	10490
5	12.07.2015	06.05.2019	29:47	1 717,395	21000	4300	13929
6	14.03.2019	06.05.2019	9:02	832,541	48000	3500	6749
7	19.03.2019	06.05.2019	9:16	831,035	39000	1900	10292
8	31.07.2018	06.05.2019	9:52	663,068	13000	1600	5282
9	20.02.2013	06.05.2019	12:52	582,082	11000	1400	9089
10	24.01.2011	06.05.2019	21:49	466,167	3900	923	7120
11	23.04.2018	06.05.2019	19:05	366,965	11000	1800	5822
12	11.12.2016	06.05.2019	6:59	295,195	9400	1100	3858
13	06.07.2012	06.05.2019	12:22	270,185	3500	1700	9057
14	03.09.2015	06.05.2019	21:55	252,367	5100	535	2368
15	12.04.2012	06.05.2019	7:06	238,094	2600	582	3303
16	05.02.2017	06.05.2019	4:44	194,203	2300	454	2672
17	04.01.2008	06.05.2019	9:59	162,475	956	142	939
18	13.07.2017	06.05.2019	51:45	152,655	2600	229	556
19	02.07.2017	06.05.2019	3:03	145,512	994	442	1287
20	05.02.2010	06.05.2019	8:14	98,266	2100	176	1218

E. Bootstrap chart

	Deniers	Neutral	Activists	NR	NO
1	0.46139	0.09027	0.27285	0.02516	0.15032
2	0.43415	0.01957	0.34834	0.0285	0.16945
3	0.43956	0.05164	0.37009	0.05954	0.07915
4	0.48971	0.02502	0.28766	0.07149	0.12612
5	0.41195	0.06968	0.39183	0.02306	0.10347
6	0.50026	0.03474	0.30098	0.06719	0.09683
7	0.37117	0.06355	0.3037	0.09054	0.17103
8	0.40985	0.059	0.28115	0.07863	0.17137
9	0.53844	0.01623	0.2531	0.01686	0.17537
10	0.33155	0.07206	0.29798	0.04872	0.24968
11	0.43626	0.05404	0.29965	0.04581	0.16425
12	0.28055	0.04337	0.33914	0.0495	0.28745
13	0.50299	0.07434	0.22307	0.05778	0.14182
14	0.45827	0.02655	0.31845	0.03114	0.16559
15	0.43681	0.04964	0.25765	0.07511	0.18079
16	0.48287	0.03506	0.31866	0.0719	0.09151
17	0.46664	0.09564	0.2533	0.04765	0.13676
18	0.56539	0.06649	0.26387	0.02277	0.08147
19	0.52154	0.0471	0.28255	0.07057	0.07825
20	0.3556	0.09532	0.29216	0.07182	0.18509
21	0.48034	0.04124	0.29007	0.06636	0.12199
22	0.50718	0.06248	0.28196	0.08969	0.05869
23	0.37603	0.14969	0.3398	0.01689	0.1176
24	0.44594	0.06979	0.279	0.08047	0.1248
25	0.35687	0.09135	0.34184	0.05985	0.15009
26	0.40699	0.08241	0.29327	0.03593	0.1814
27	0.52453	0	0.32877	0	0.1467
28	0.474	0.0643	0.20761	0.08608	0.16801
29	0.42442	0.06625	0.33831	0.05281	0.11821
30	0.44755	0.09765	0.27332	0.05142	0.13006
31	0.51566	0.026	0.28722	0.046	0.12512
32	0.50601	0.00856	0.33682	0.04003	0.10859
33	0.53169	0.02868	0.28934	0.04919	0.1011
34	0.52819	0.01305	0.28948	0.03361	0.13567
35	0.45934	0.09146	0.2781	0.02691	0.14419
36	0.45831	0.05953	0.31169	0.03761	0.13286
37	0.2679	0.02601	0.45386	0	0.25223
38	0.39868	0.11438	0.18896	0.02665	0.27133
39	0.30276	0.09611	0.39984	0.08657	0.11471
40	0.42867	0.06074	0.21806	0.06836	0.22417
41	0.39593	0.08927	0.3409	0.02475	0.14916
42	0.5016	0.03447	0.22466	0.05493	0.18435
43	0.41105	0.0531	0.27673	0.04765	0.21146
44	0.32514	0.05125	0.37027	0.09484	0.1585
45	0.50974	0.00551	0.36928	0.02563	0.08985
46	0.43208	0.07212	0.26602	0.07201	0.15777
47	0.51268	0.03418	0.32444	0.01549	0.11321
48	0.40825	0.0533	0.26687	0.06264	0.20893
49	0.47533	0.06479	0.31746	0.0079	0.13451
50	0.43585	0.04783	0.35118	0.05066	0.11448

	Deniers	Neutral	Activists	KO	NR
51	0.437976	0.069989	0.36404	0.057762	0.070232
52	0.539529	0.034532	0.218236	0.059878	0.147825
53	0.436724	0.067774	0.376321	0.046276	0.072906
54	0.446224	0.087744	0.372284	0.053452	0.040296
55	0.513835	0.024014	0.318684	0.011108	0.13236
56	0.468689	0.047148	0.331696	0.030014	0.122454
57	0.402878	0.07006	0.277947	0.068685	0.180431
58	0.431876	0.053052	0.317662	0.067884	0.129525
59	0.494424	0.034339	0.281647	0.040602	0.148987
60	0.456623	0.092582	0.194297	0.075921	0.180577
61	0.434242	0.027871	0.305311	0.092642	0.139933
62	0.440694	0.039209	0.274769	0.082445	0.162882
63	0.418184	0.072192	0.269116	0.088866	0.151642
64	0.411071	0.039385	0.331129	0.069702	0.148712
65	0.410027	0.036212	0.369825	0.037793	0.146142
66	0.323109	0.101961	0.351397	0.109109	0.114424
67	0.439231	0.004164	0.362071	0.050073	0.14446
68	0.372055	0.064337	0.328456	0.044713	0.190439
69	0.390731	0.035393	0.262257	0.101849	0.20977
70	0.435052	0	0.343143	0.07025	0.151556
71	0.467013	0.018848	0.324368	0.044766	0.145006
72	0.374251	0.02391	0.350811	0.059395	0.191633
73	0.410009	0.050231	0.279979	0.077338	0.182444
74	0.405377	0.042703	0.33692	0.067038	0.147962
75	0.36584	0.059392	0.235944	0.111701	0.227124
76	0.396369	0.071631	0.386591	0.046942	0.098466
77	0.326971	0.042738	0.295866	0.097301	0.237124
78	0.312456	0.021781	0.393175	0.057693	0.214894
79	0.384499	0.098779	0.338121	0.050959	0.127643
80	0.484819	0.059577	0.315883	0.033484	0.106236
81	0.503401	0.069235	0.230388	0.03296	0.164017
82	0.425811	0.022645	0.366933	0.059148	0.125462
83	0.574635	0.014186	0.329266	0.006421	0.075492
84	0.464804	0.053072	0.275333	0.057217	0.149575
85	0.430107	0.084996	0.325283	0.055648	0.103966
86	0.510022	0.019344	0.330899	0.064374	0.07536
87	0.462842	0.038139	0.341698	0.06205	0.095272
88	0.401315	0.109131	0.231804	0.07554	0.182209
89	0.421636	0.065181	0.301617	0.023707	0.18786
90	0.403688	0.046375	0.414326	0.011508	0.124103
91	0.473522	0	0.259418	0.029456	0.237605
92	0.586221	0.031149	0.254142	0.022342	0.106146
93	0.428185	0.011004	0.320328	0.059886	0.180597
94	0.496072	0.080239	0.296872	0.019017	0.1078
95	0.483677	0.080389	0.230505	0.051166	0.154263
96	0.40952	0.080049	0.264597	0.104551	0.141284
97	0.432282	0.042845	0.320494	0.044952	0.159426
98	0.398957	0.015605	0.300926	0.080378	0.204134
99	0.485862	0.062327	0.261166	0.03003	0.160615
100	0.457586	0.073287	0.290883	0.051712	0.126532