LUND UNIVERSITY
School of Economics and Management

Exploring the data behind students’ published theses

- *Analyzing the pride of Lund University*

Bachelor Thesis in Statistics, STAH11 15 hp
Lund University School of Economics and Management

Author:
Per Granberg

Supervisor:
Björn Holmquist

Abstract

Lund University have for over ten years been using a website called LUP Student Papers where they publish theses from bachelor and master courses. The aim of this thesis is to use visualizations and data mining techniques that will explore and shows interesting aspects of the data. The wide variety of variables in the dataset can be used to gain insight regarding several interesting questions such as are the number of theses increasing for each year? Is a thesis in English becoming more common? How many times have a thesis been downloaded on average?

The second purpose of this thesis is to use a Random Forest model and classify the abstract into three faculties, LUSEM, Social Science and Engineering. The aim is to see if the three faculties can be easily classified which would suggest that there is some noticeable difference in the text between the faculties. The abstracts had to be preprocessed with natural language processing techniques such as tokenization and stemming. The classification model achieved a relatively good accuracy around 0.80 and therefore suggest that the abstract can be classified. Further research can focus on different models for text classification.

Keywords*: data visualization, Lund University, classification model, NLP.*

Table of Contents

1. Introduction 1

1.1. Research purpose 2

1.2. Thesis structure 2

2. Data results and visualization 3

2.1. Obtaining the data 3

2.2 Visualizing Missing Values 4

2.3. Student papers across the years 5

2.3.1. Different paper types over the years 6

2.3.2. Visualization of Language across the years 8

2.4. Faculties 10

2.4.1 Treeplot over Faculties and paper Type 11

2.5. Visualizing the Keywords 12

2.5.1. Wordcloud over the most common keywords 12

2.5.2. Sentimental analysis 13

2.6. Analyzing gender trends 14

2.6.1. Published student papers based on gender 14

26.2. Does people tend to cowrite with the same gender? 15

2.6.3. Gender across the faculties 16

2.7. Analyzing downloads 17

2.7.1. Faculties and the average download 18

3. Classifying Faculties based on abstract 19

3.1. Cleaning the text data 20

3.2 Random Forest as the Classification Model 23

4. Empirical Findings and Analysis 24

4.1. Variable importance 24

4.2. Confusion matrix and statistics over model performance 25

4.3. Summary statistics over k-fold cross-validation 26

5. Conclusion 28

5.1. Summary 28

5.2. Further research 29

References: 30

Appendix A – LUP website information 32

Appendix B – Packages used in R 34

Appendix C – 15 Random Names and gender 35

Index of Tables and Figures

Tables:

Table 1: Overview for the variables obtained from web-scraping LUP 3

Table 2: Total number of papers in English 19

Table 3: Number of abstracts used in the model 19

Table 4: Example of the document-term matrix 20

Table 5: Summary over the ten most common words for each faculty (pre stemming) 21

Table 6: Confusion matrix for the classification model; Accuracy = 0.8022 25

Table 7: Summary over model statistics for the faculties 26

Figures:

Figure 1: A graph that provides information on which observation that has missing values 4

Figure 2: Number of published papers over all the years 5

Figure 3: The paper types over the years 6

Figure 4: Paper type in percentage across the years 7

Figure 5: The amount of papers published in Swedish or English over the years 8

Figure 6: The use of Swedish and English across the different Degree types 9

Figure 7: Student papers ordered by each Faculty 10

Figure 8: Faculties and the Type of student papers 11

Figure 9: Word-cloud over the most frequently used Keywords 12

Figure 10: Sentimental analysis over the keywords 13

Figure 11: Female and Male published papers across the years 14

Figure 12: Two plots that investigate if two authors often have the same gender 15

Figure 13: Female and Male published papers across the years 16

Figure 14: Histogram over total downloads 17

Figure 15: Boxplot over Faculties and their average downloads 18

Figure 16: TF-IDF diagram over 15 words for each faculty 22

Figure 17: Variable importance from the RF-model 24

Figure 18: Example of k-fold cross-validation where k = 5 26

1. Introduction

The information on the internet is steadily growing for each day. Webpages that contain information can be utilized and downloaded by using web-scraping tools, these tools can save the data from webpages in a format that can be used for analysis (Vargiu & Urru, 2013).

Lund University uses the webpage LUP Student Papers (LUP) in order to publish all the bachelor and master theses from the students. The webpage contains information such as the authors name, the title, how often the thesis has been downloaded and much more (see appendix A). Lund University have been using LUP for over ten years so the webpage contains a lot of information that could be interesting to analyze and draw conclusions from.

Data mining is a concept derived from exploring a dataset in order to create a better understanding or find interesting attributes that can be hidden in the data. It has grown rapidly in popularity and several methods has been developed and further improved thanks to computer power and new technology (Zaki & Meira, 2014).

A critical role for a statistician is to visualize and show important aspects that can be explained by observing the data (James, Witten, Hastie, & Tibshirani, 2013). Visualizations are necessary for conveying information to individuals and are often used in newspapers, research papers and annual reports as an effective way to showcase the underlaying data (Healy, 2019).

Data visualization and storytelling are often ranked as one of the most important skills needed for data scientist and statisticians. The reason for this is because data visualizations often requires cleaning the data and selecting the wanted properties, for example, grouping the variables, calculating the mean for several groups, etcetera (Healy, 2019).

Several tools and software have become more focused on displaying data, some of the most popular software are R, Python, Power BI and Tableau. R and Python can also be used to create statistical models, they have therefore become the most common options to use for data scientists (Zaki & Meira, 2014).

Text analysis and Natural Language Processing (NLP) is a growing subgroup in data mining and the rise in popularity is due to the fact that more text data is available and the techniques are able to better analyze and compute very large amount of text (Ledolter, 2013). NLP is often used when evaluating the sentiment of a text block, for example if a tweet has a negative or positive attitude or classifying text into groups.

This thesis will focus on data that is web-scraped from LUP and will therefore investigate and use data visualization to explore aspects regarding the student theses published at Lund University. Each thesis on LUP has written an abstract that provides a short description of the thesis, this text will be analyzed using statistical methods in order to classify the thesis into three different groups that represent the faculty that the thesis belongs to (LUSEM, Social Science or Engineering, see chapter 3).

1.1. Research purpose

The large amount of data that was obtained from LUP will be used for two purposes:

1: Create visualizations that showcases interesting aspects in the dataset.

2: Use statistical approaches and create a statistical model that will see if it is possible to obtain a high text classification accuracy concerning the abstract of the thesis and which faculty it belongs to.

1.2. Thesis structure

The structure of the thesis is as follows: the first chapter starts with an introduction to the field of web-scraping and how important data visualization is. The second chapter provides a short summary of the variables that was obtained from web-scraping LUP. Chapter 2 also shows a lot of visualizations and explain interesting aspect of the data in order to complete the first purpose. The method used for cleaning the data and prepare it for text classification is explained in the third chapter. The result of the classification model and summary statistics such as accuracy, sensitivity and specificity are presented in chapter 4. Finally, the last chapter provides a summary of the paper as well as suggestions for further research.

2. Data results and visualization

The first purpose of this study is to create visualizations of the dataset and generate insights regarding student papers published at Lund University. This chapter will therefore focus heavily on figures and text that will explain what the graphs indicates. The approach will mimic how data scientists often work with a new dataset, which is to create graphs in order to get a better understanding of the data (Ledolter, 2013).

All the visualizations are made in R using packages such as ggplot2 and its extensions.

2.1. Obtaining the data

R is a multifunction tool that by the help of packages can do much more than just statistics. The package Rvest is the package that was used for web-scraping.

The data consists of 62 865 observations (unique student papers) and 18 variables. Most of the variables are factors such as "*Course", "Type"* and *"Supervisor"* but variables like *"Year"* and how many times the paper have been downloaded are integers. Table 1 below displays the variables that was scraped from LUP:

Table 1: Overview for the variables obtained from web-scraping LUP, 62 865 observations and 18 variables

Variable	Type	Description
Title	Character	The title for the paper.
Names	Character	The names of the authors.
Abstract	Character	The abstract of the thesis.
AbstractEng	Character	The abstract of the thesis in English.
Supervisor	Character	The supervisor(s) for the paper.
Department	Character	The department, example "Department of Statistic" etc.
Course	Character	The name of the course.
Year	Integer	The year the paper was published; "2019", "2018" etc.
Type	Character	Type of paper, "Bachelor", "Master one-year" etc.
Subject	Character	The subject of the paper, example "Mathematics and Statistics".
Keywords	Character	The keywords for the thesis, around five keywords for every paper.
Language	Character	The language the paper is written in, "Swedish", "English" etc.
Open	Character	If the paper is open access and can be downloaded; "Yes" or "No".
id	Integer	The unique id number for each thesis used in web-scraping.
numauthors	Integer	How many authors that wrote the paper; "1", "2" etc.
Facaulty	Character	The faculty that the thesis belongs to.
Total download	Integer	How many times the paper has been downloaded.
Download_year	Integer	How many times the paper has been downloaded for each year.

2.2 Visualizing Missing Values

When analyzing data one important feature to first consider is the missing values (James et al, 2013; Ledolter, 2013) If the data consists of missing values one can use several methods to counter this, for example using the rest of the data in order to create data points for the missing values. Another approach is to simply remove the observations that contains missing values (James et al, 2013).

The missing values in this thesis will simply not be shown in the visualizations. This is because a category for the missing values will only make the graphs harder to interpret and will not add much insight in this case.

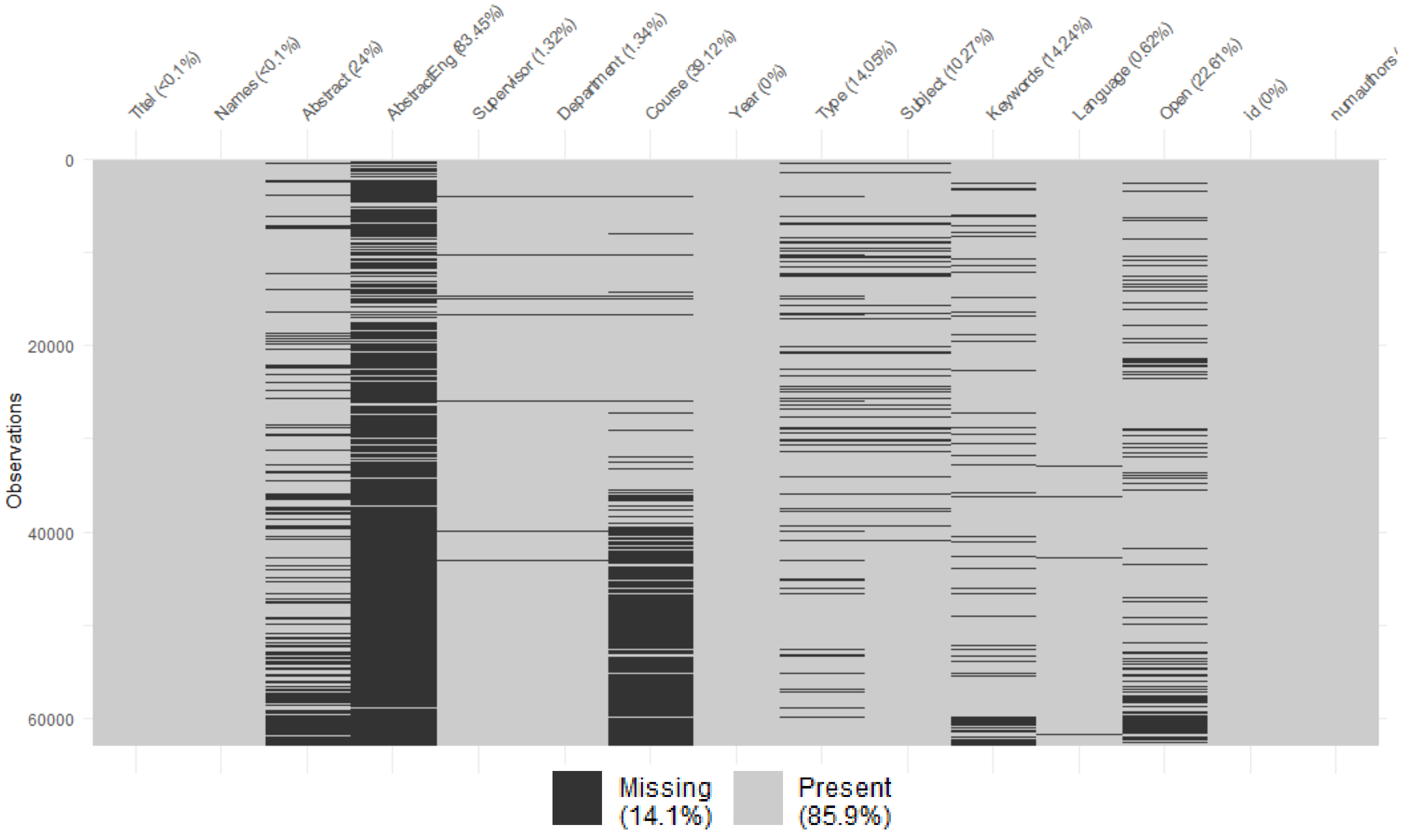


Figure 1: A graph that provides information on which observation that has missing values

The figure above shows the whole dataset and makes in clear how many missing values there are, a total of 14 percent of the variables are missing.

The variable *"AbstractEng"* have the highest score of missing observations with 83.45 %, this is due to the fact that most students that write in English only post one abstract and that is inside the regular *"Abstract"* variable.

The graph is ordered by year, meaning that the newest papers are at the top and the oldest papers are at the bottom. When the paper was published is clearly a factor when viewing the missing values for the variable *"Course"*, the missing values almost only exist on the lower half, indicating that older students did not add the course when publishing their thesis.

The only real problem with the missing values in this thesis is the loss of information concerning abstracts written in English (see section 3).

2.3. Student papers across the years

One of the first question that we can deduce from the data is; *are the number of published student papers rising, declining or on the same level for every year*? The students at Swedish Universities has increased for several years (SCB, 2020), it would therefore be expected that the number of published papers at Lund are also rising as a reaction to more students.

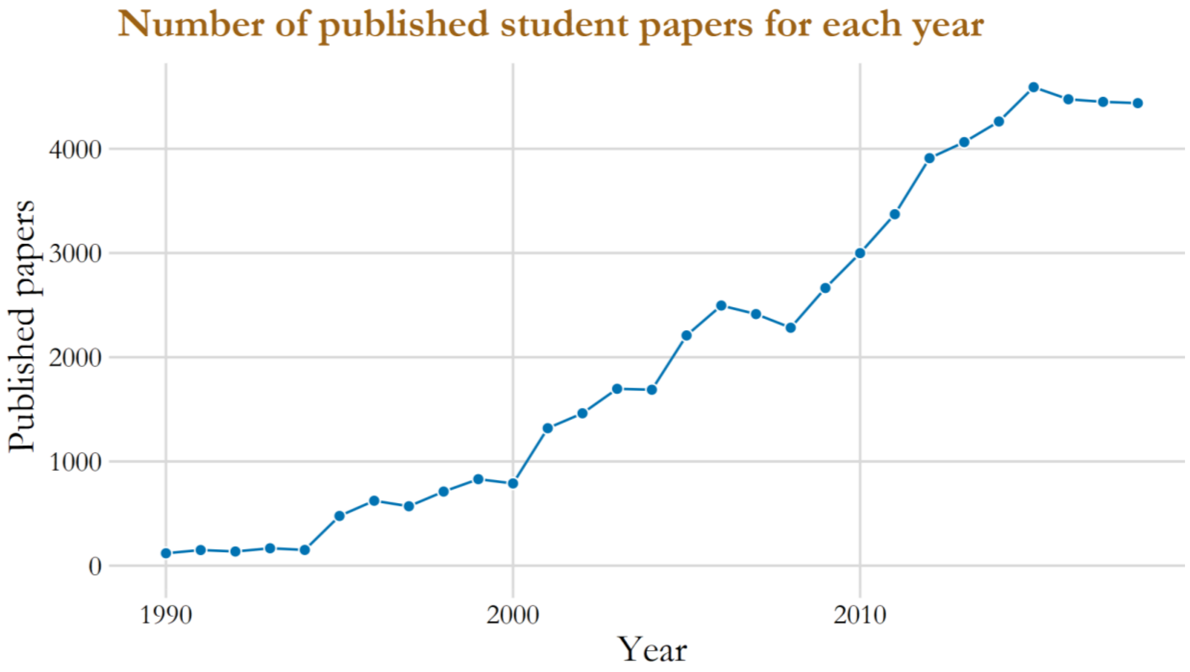


Figure 2: Number of published papers over all the years

The graph above provides the insight that the number of student papers have grown for almost every year, the sharpest rise started around the year 2008 and lasted to 2015. The yearly number of papers seems to have been stabilized in the recent years with around 4 500 papers published.

2.3.1. Different paper types over the years

The variable *"Type"* consists of factors that tells if the paper is a bachelor, master (one year) or master (two years) degree. The data have more levels (*Phd*, *Licent* etc.) but this thesis will only focus on the three mentioned above since they are the most common.

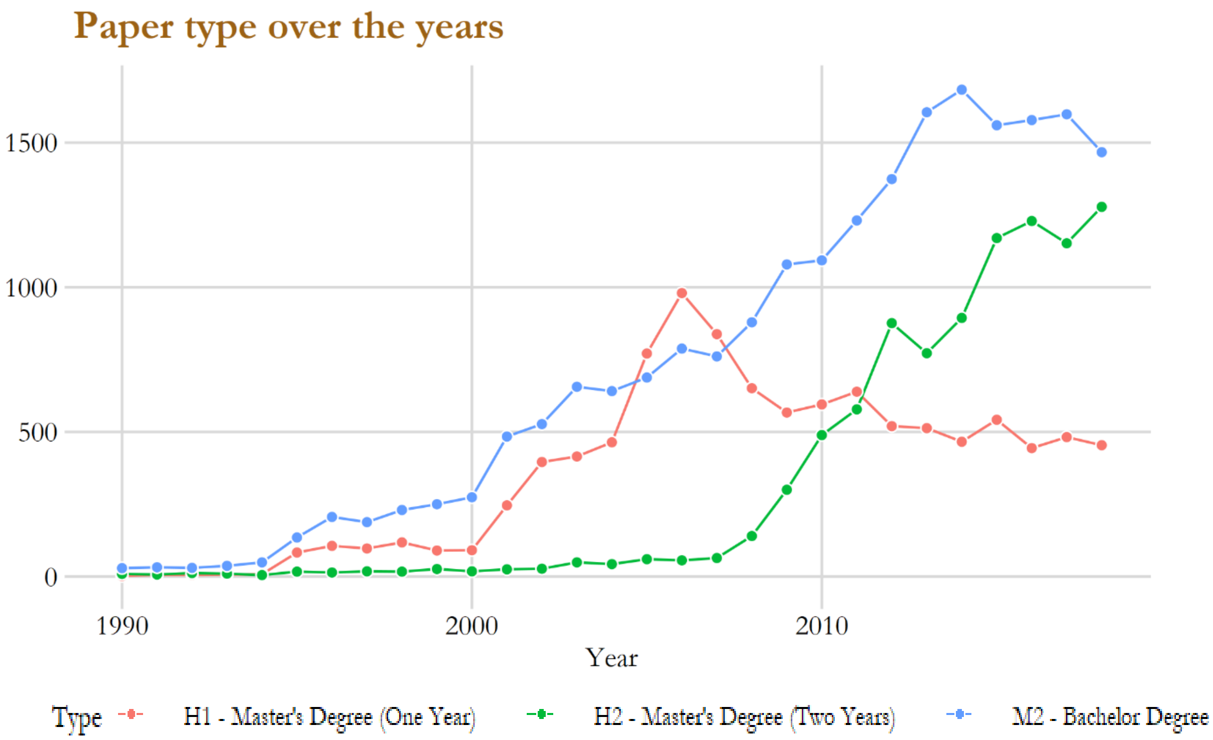


Figure 3: The paper types over the years

The graph above illustrates that papers made in the Bachelor level are most common, except for three years between 2005 to 2007 where the most common paper was master's degree (One Year).

One of the most interesting things to notice in the graph is that master's degree (Two Years) started to get momentum around the year 2008 and have increased ever since. On the other hand, master's degree (One Year) decreased for every year after the year 2006 and became even lower than master's degree (Two Year) in the year 2012.

On the next page is a percentage graph presented that visualize more effectively the allocation of the paper types across the years.

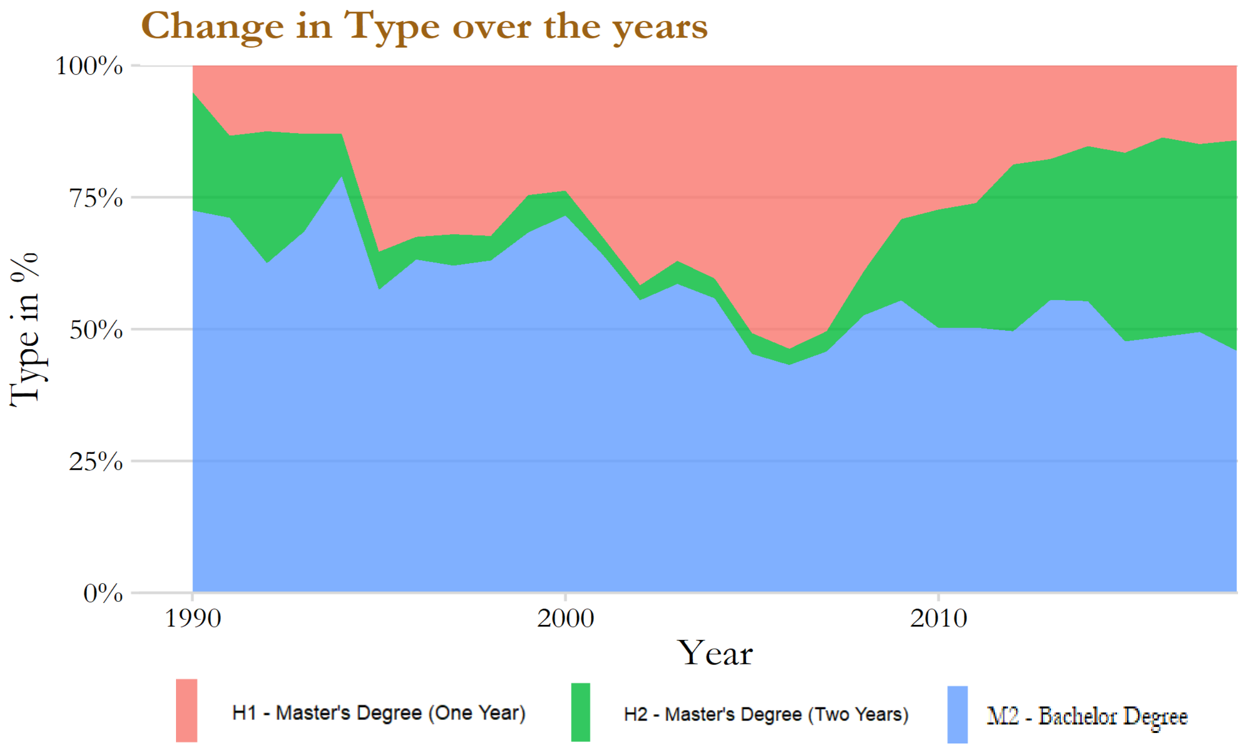


Figure 4: Paper type in percentage across the years

The percentage plot indicates that in the year 2018 that almost 50 percent of all the papers published are a bachelor’s degree, while around 30 percent are master’s degree (Two Year) and roughly around 20 percent are master’s degree (One Year).

2.3.2. Visualization of Language across the years

The data consists of a total of four languages, Swedish, English, Danish and French. This thesis will only focus on Swedish and English because the rest had very few observations and would only distort the visualizations and insights.

Let’s first answer the question, *how has the use of Swedish and English changed over time*?

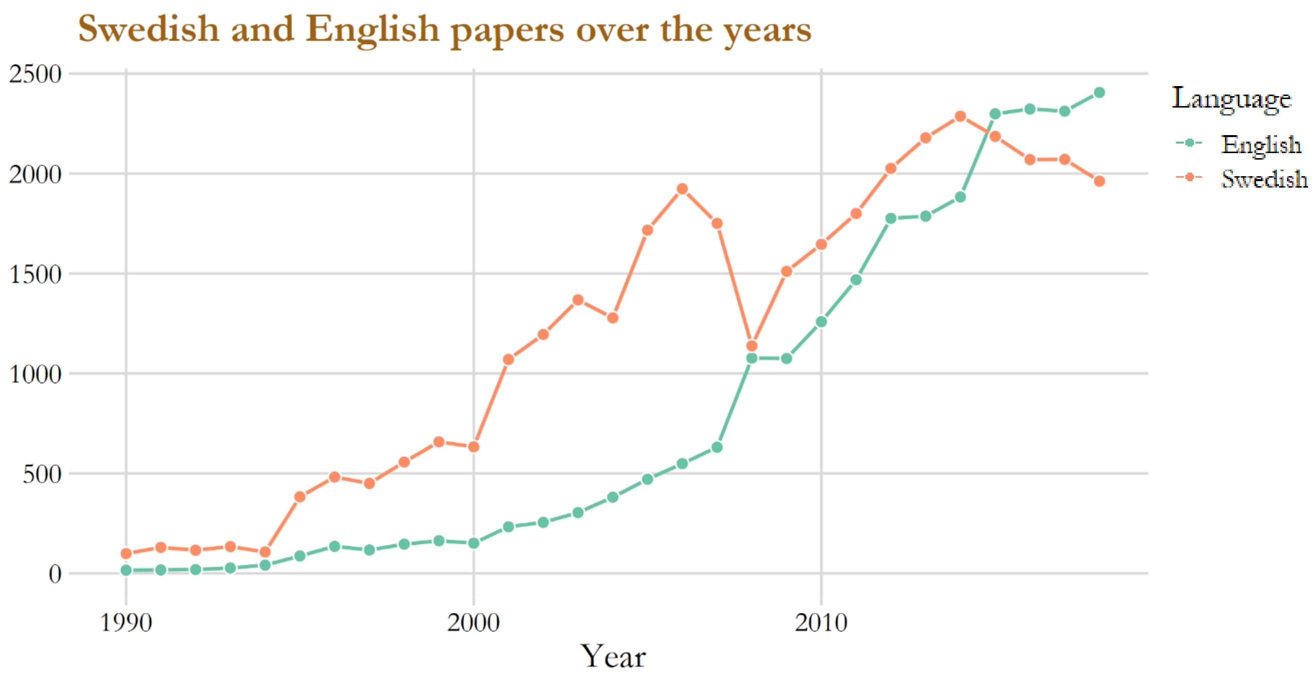


Figure 5: The amount of papers published in Swedish or English over the years

Swedish has been the most popular language from 1990 to 2015. The biggest change happened around the year 2008 where there was a sudden decrease in the use of Swedish and an increase in English.

The number of Swedish student papers has slightly decreased over the last five year while English has instead increased. This might be because universities are becoming more international and writing in English is more meritorious (Björkman, 2008).

Education at the higher levels are usually more international and English is often the language that is used, it is therefore interesting to see if the data support the belief that the use of English is more prominent the more advanced the education is (master levels against bachelor).

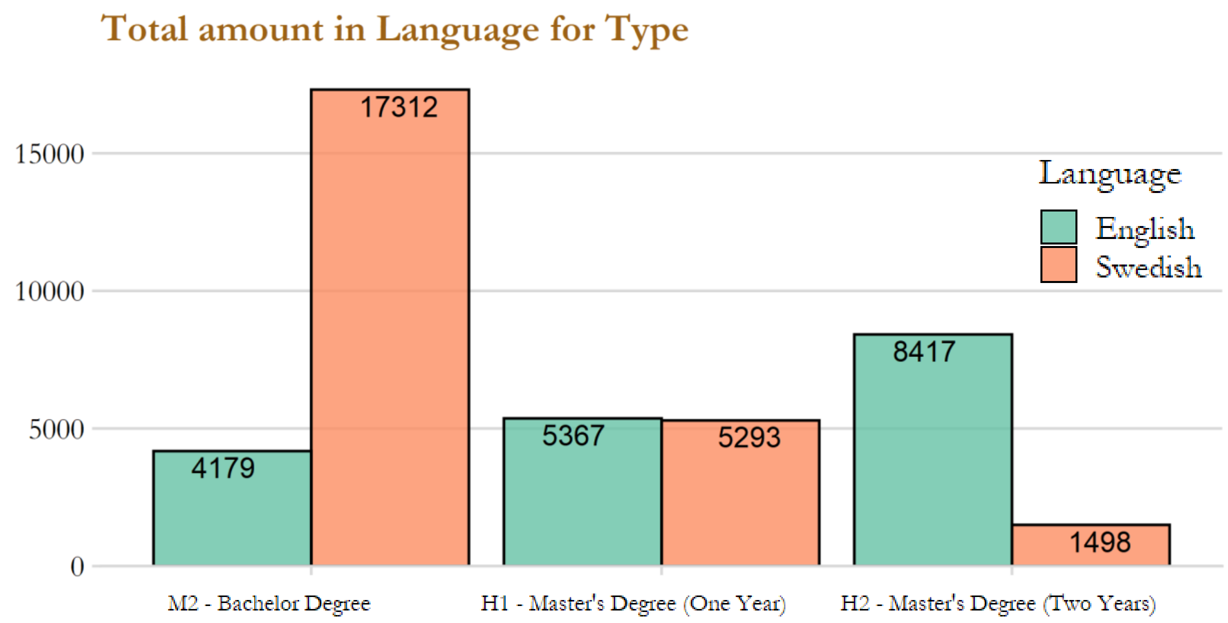


Figure 6: The use of Swedish and English across the different Degree types

It is clear that Swedish is most used at the bachelor level and English is instead highly common in master’s degree (Two Years).

The result is not surprising since with higher and more advanced courses are students expected to perform better, and one aspect of this is often to have courses in English since that is the common research language of the world (Björkman, 2008).

2.4. Faculties

Lund University consist of six faculties, such as LUSEM (Lund University School of Economics and Management), Social Sciences, Law and more. It is therefore interesting to investigate if each faculty has produced the same amount of student papers.

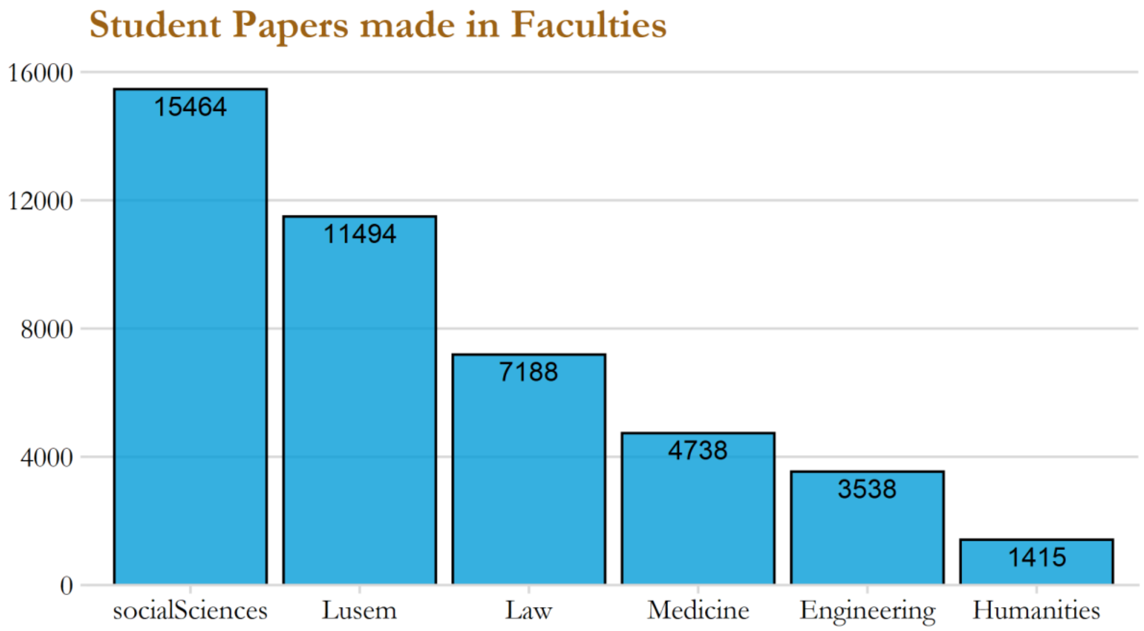


Figure 7: Student papers ordered by each Faculty

We can tell from the graph that Social Sciences and LUSEM are the two faculties that have produced highest quantity of student papers. The faculties Engineering and Humanities are instead at the bottom, indicating that these faculties have less students.

The variable *"Type"* creates a little bit of insight to why Social Sciences and LUSEM might be the best producing faculties concerning the amount of papers. These two faculties are namely also the ones that have published the most bachelor's degrees, see figure 8 on the next page.

2.4.1 Treeplot over Faculties and paper Type

Another relationship that is interesting to investigate is if the faculties have the same allocation between each other considering how many bachelors or master degree papers they produce.

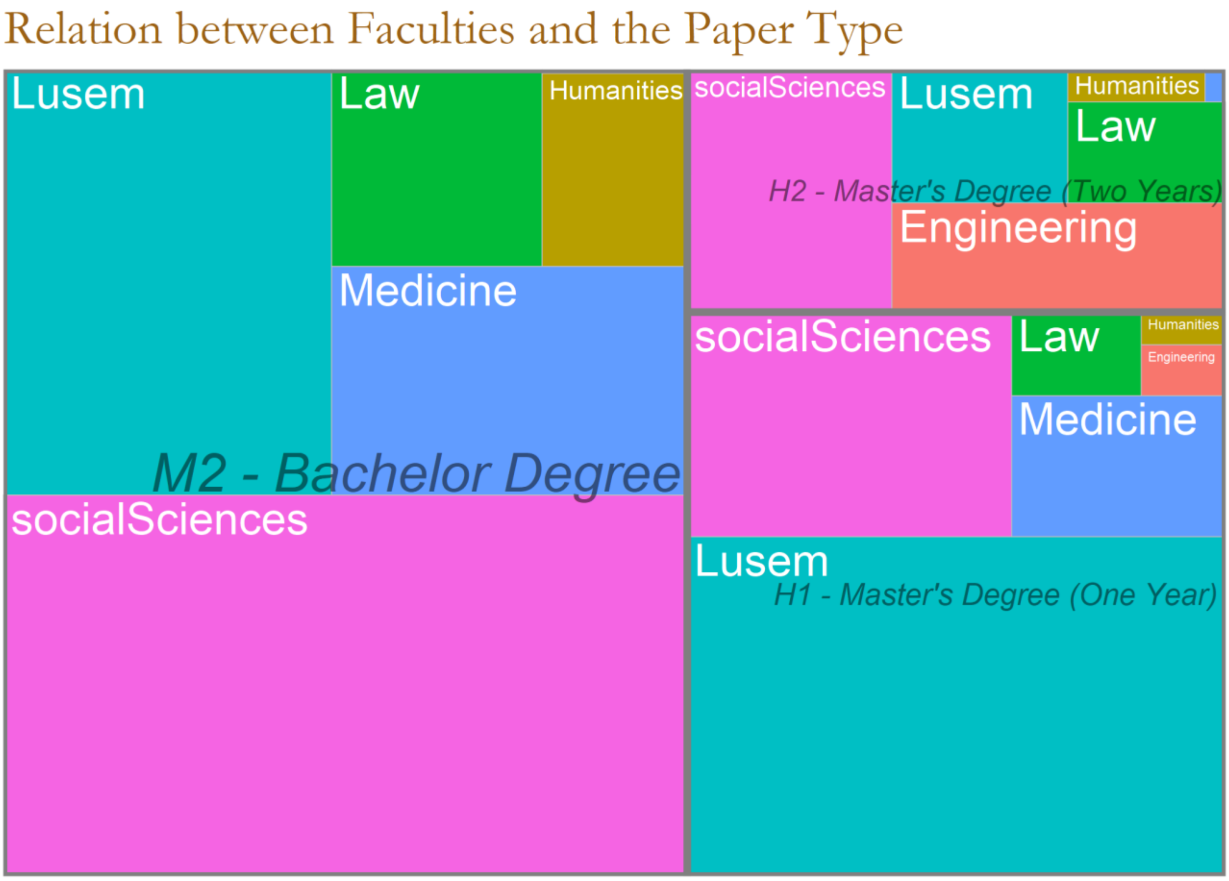


Figure 8: Faculties and the Type of student papers

The treeplot shows that the left side that consist of bachelor’s degree cover around half the plot, indicating that around 50 percent of the student papers are bachelor’s degrees, which is consistent with figure 4.

The figure also shows that engineering almost only publishes papers that are two years master’s degree.

2.5. Visualizing the Keywords

Students are encouraged to add keywords when publishing their thesis, so it is easy for others to find it on the internet. This section will focus on displaying the most common keywords and use NLP in order so select the most negative and positive keywords.

2.5.1. Wordcloud over the most common keywords

The wordcloud plot can be used to investigate the most occurred words in a text (Ledolter, 2013). The positive aspect of the plot is that it can contain a lot of words and the frequency of a word makes it bigger. The negative side is that the words are not ordered, and it becomes unclear regarding the number of times the specific word appears (Healy, 2019).

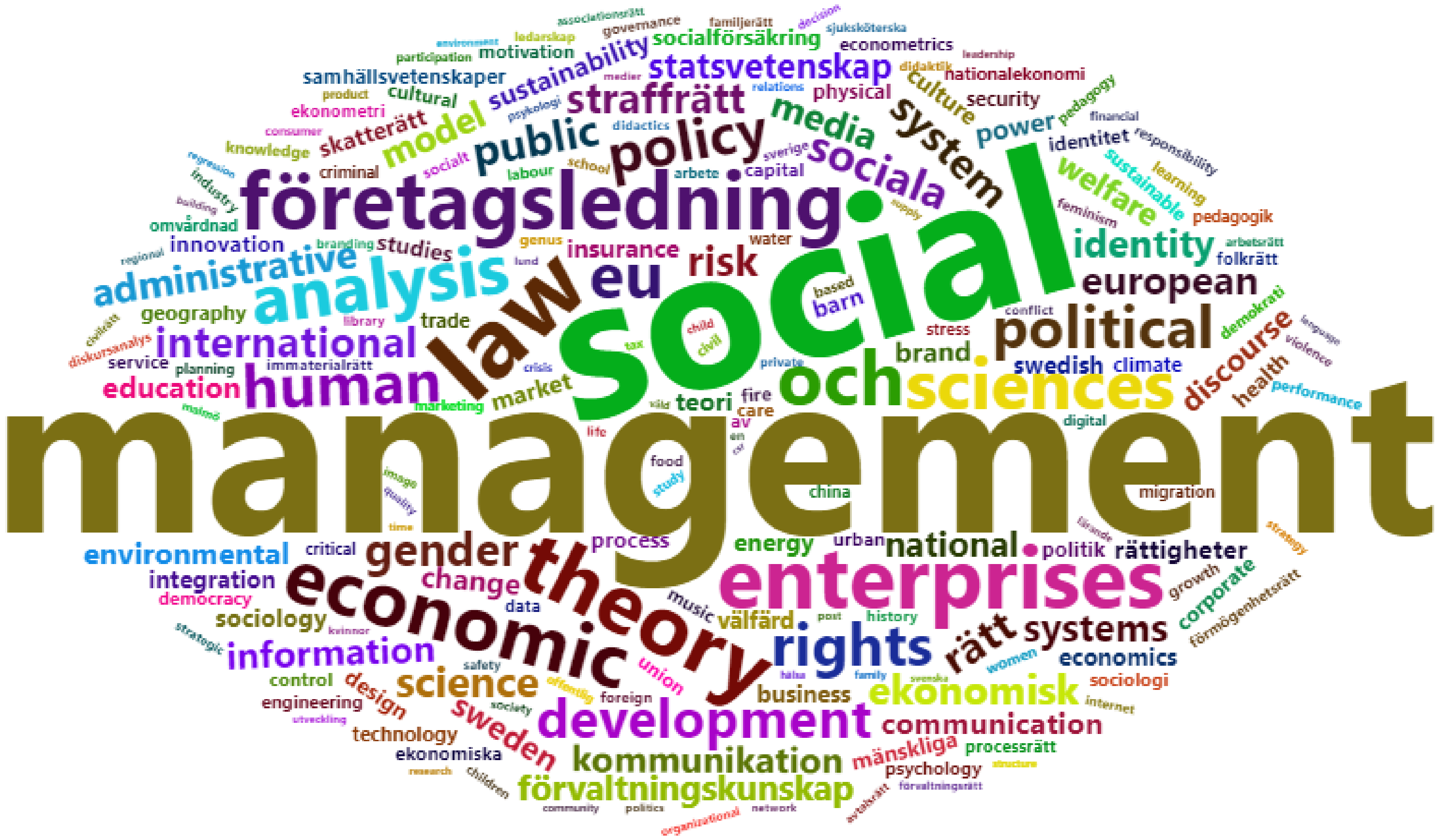


Figure 9: Word-cloud over the most frequently used Keywords

Some of the biggest word in the plot are: *Management, Social, Enterprises, Företagsledning, Law and Theory*. This indicates that the most common words belong to the economic and social studies, which makes sense considering that the faculties Social Science and LUSEM produced the most student papers (see figure 7).

2.5.2. Sentimental analysis

There are several different approaches to conduct a sentimental analysis by using NLP. The approach this thesis will apply is to simply classify which word that have a negative or positive sentiment. The method is conducted by finding a match with words from a database that also indicates if the word har a positive or negative meaning (Manning & Schütze, 1999).

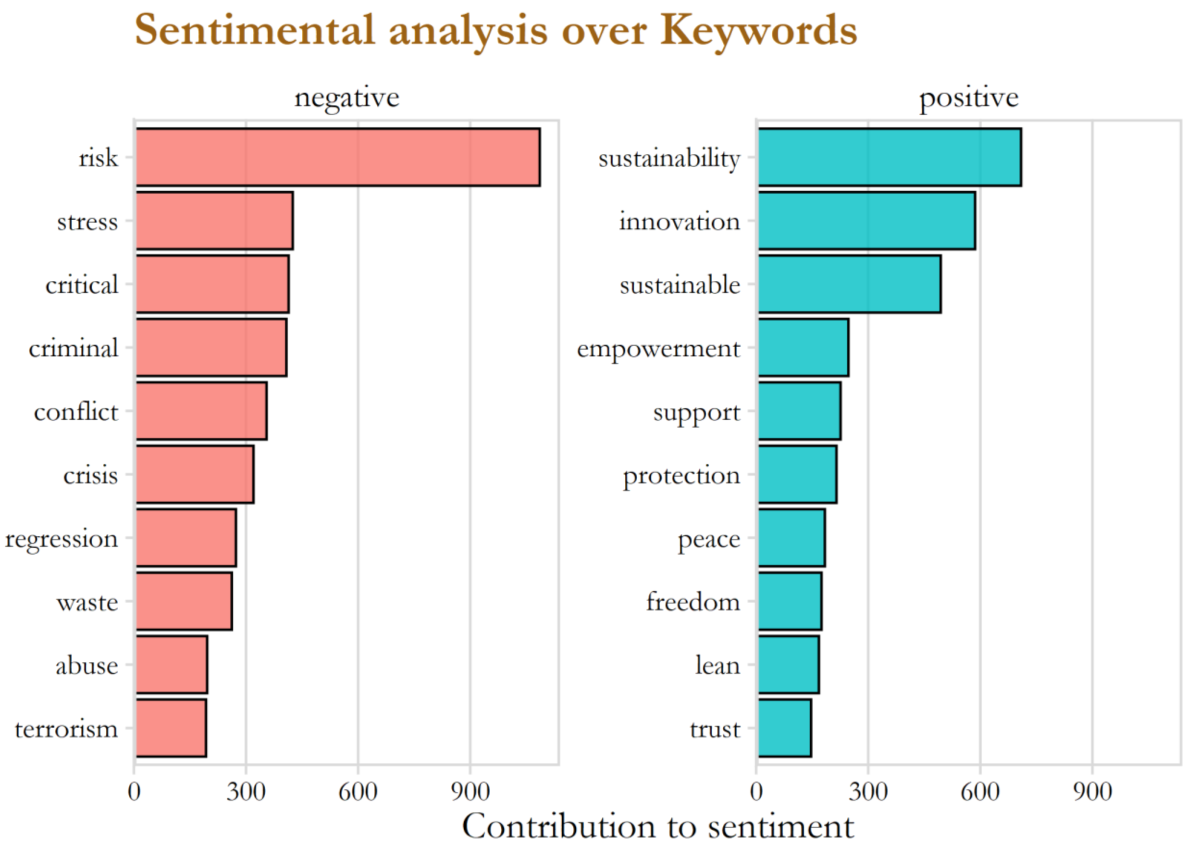


Figure 10: Sentimental analysis over the keywords

The graph above shows the ten most frequent positive and negative keywords that was able to match the words from the database.

The most common negative words are *risk*, *stress*, *critical* and *criminal*. However, the word *regression* is classified as negative but there is a high chance that it simply is describing the statistical model linear regression and is therefore not a true negative word in that context.

Sustainability, *innovation*, *sustainable* and *empowerment* are the most occurred positive keywords.

2.6. Analyzing gender trends

The raw data scraped from the LUP website contains only the names of the authors. However, by extracting the first name of each author it is possible to use the R-package Genderize to determine if the name is associated with females or males. The package is not perfect so when it does not know the gender of a name it will return a missing value, which is better than random guessing because then it is at least possible to sort out the missing values. See appendix C that show 15 randomly selected first names and their assigned gender according to Genderize. I assume that the percentage of misclassifications are the same for males and females.

2.6.1. Published student papers based on gender

According to SCB (2020) so are there more females attending universities than there are males. It's possible to see if that statement is true in this dataset by observing the trends of published papers for each gender.

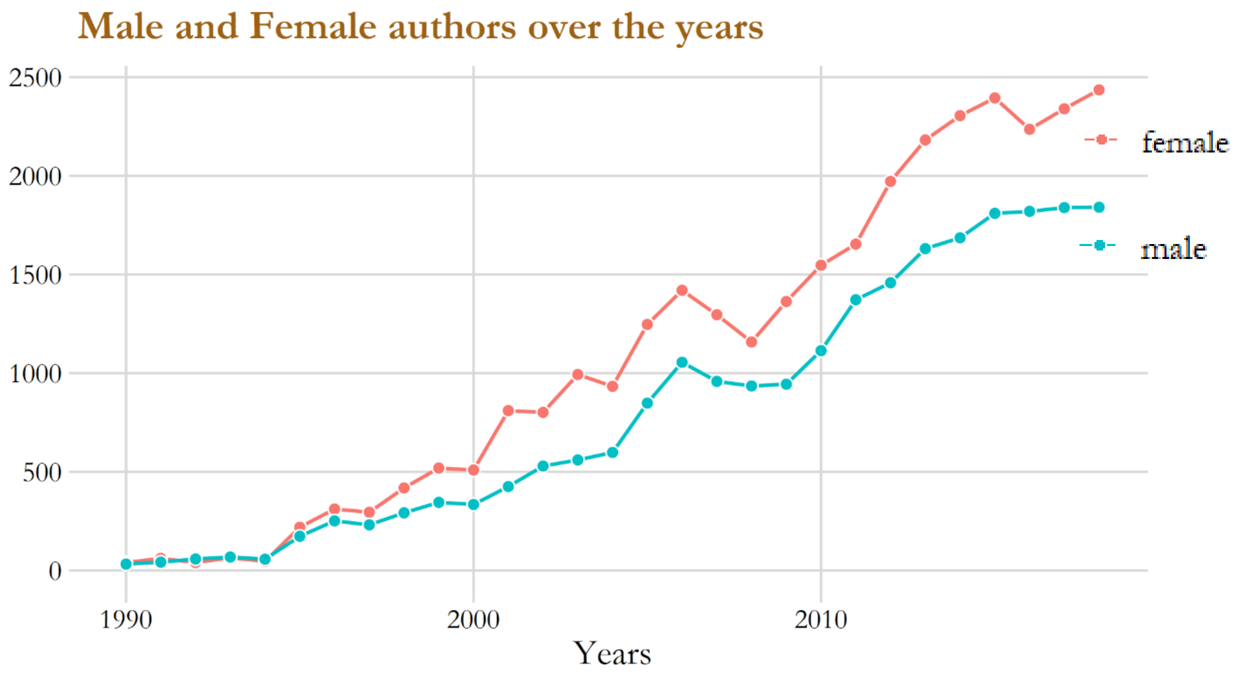


Figure 11: Female and Male published papers across the years

The above figure agrees with SCB (2020) since females have always published more papers, indicating that there are more females at Lund University.

26.2. Does people tend to cowrite with the same gender?

An interesting question to ask is if students are more likely to work together with another one that have the same gender. Cinamon & Rich (2002) states that humans are more prone to work together with persons of the same gender since it increases self-esteem and makes the teamwork easier since humans are often accustomed with the same gender.

I will use the data by only considering papers that are made by two authors, I will then check if the thesis is written by two females, two males or one female and one male.

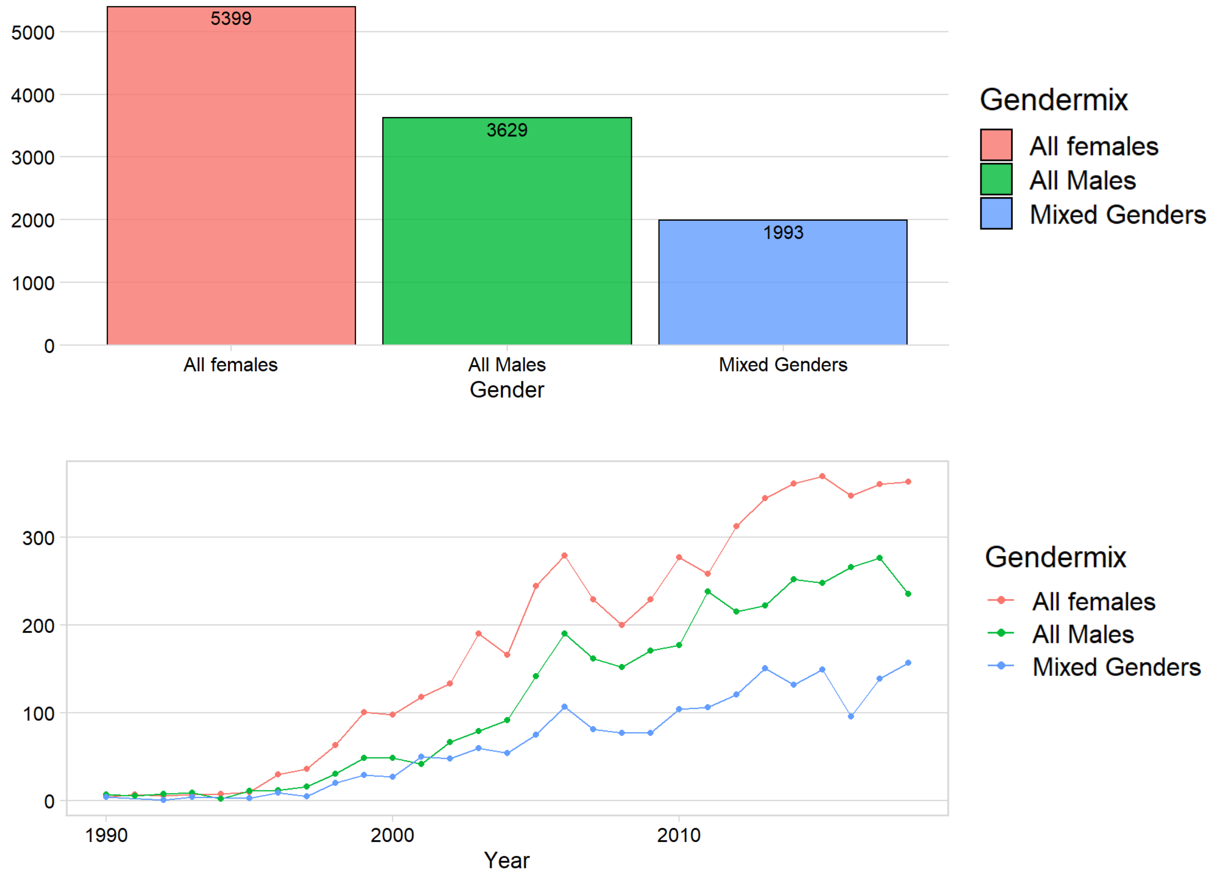


Figure 12: Two plots that investigate if two authors often have the same gender

The above plot tells us that most thesis are written within the same genders, only around 22 % of the thesis seems to be written by a male and a female. The below plot that have time on the x-axis shows that the difference seems to be constant across time, except for the two last year's where mixed genders have increased compared to the two other groups.

2.6.3. Gender across the faculties

Schilt & Wiswall (2008) writes that females tend to work in professions such as nursery, human resources and jobs with a lot of social interactions while males on the other hand works in professions like lawyer, businessman and programming. It is therefore interesting to see if the data shows us a big difference between the genders concerning which faculties they are in.

By observing how many papers are published in each faculty and if the authors are a male or female makes it possible to calculate the difference between the genders and therefore determine if the conclusions from Schilt & Wiswall (2008) holds true in this case.

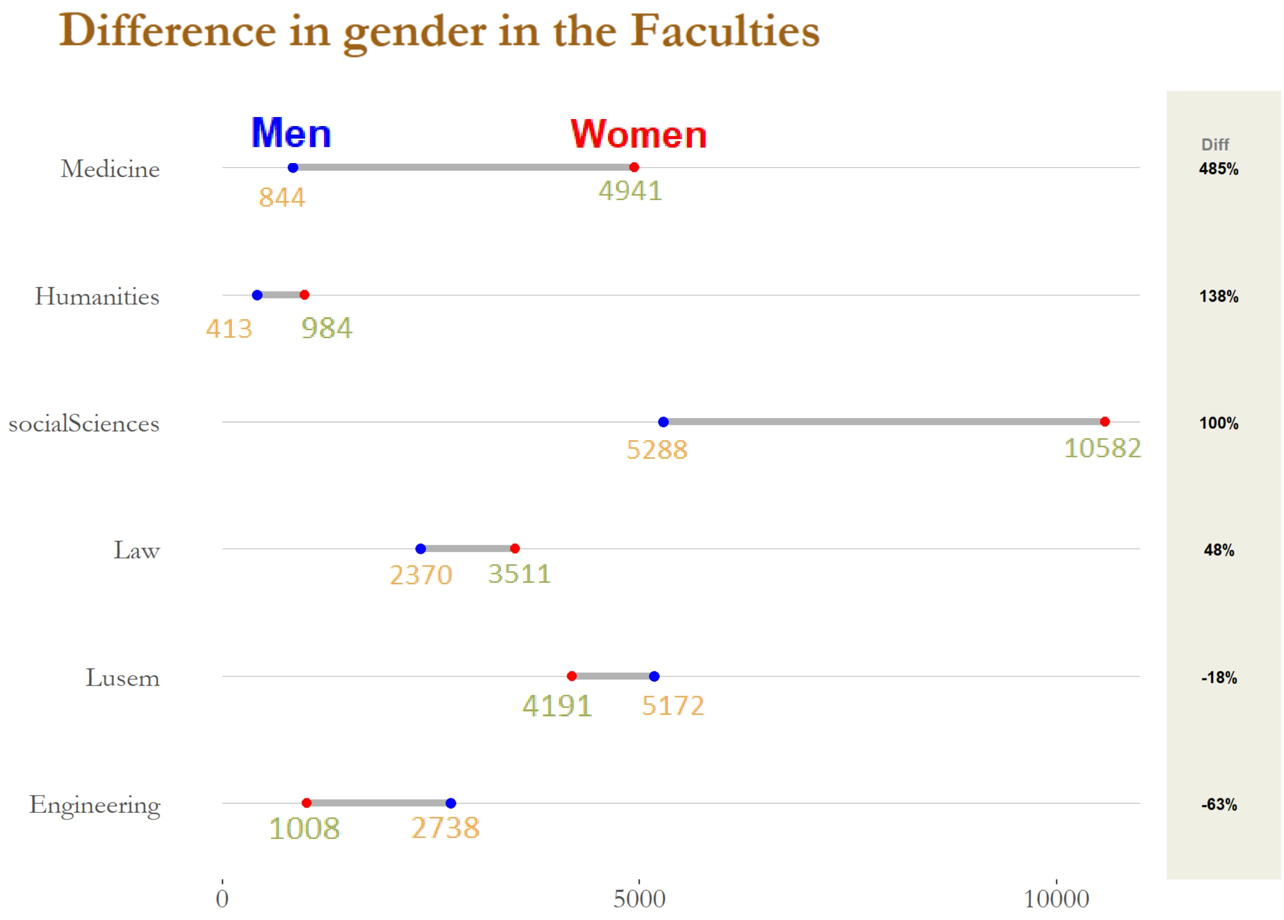


Figure 13: Female and Male published papers across the years

The above graph shows that there are more females on four out of the six selected faculties. The biggest difference in percentage is between Medicine where there are 485 % more females according to the data. The two faculties that have more males are LUSEM which is mostly programs in business and economics and the Engineering faculty. The conclusion is therefore that Schilt & Wiswall (2008) seems to be correct, since the only faculties that have more males are the ones focused on economy and technology.

2.7. Analyzing downloads

The student papers that are available to read online also contains data regarding how many times the paper have been downloaded, what country that downloaded the paper and history of downloaded for the previously 12 months.

Let's start by examining the distribution of downloads by creating a histogram.

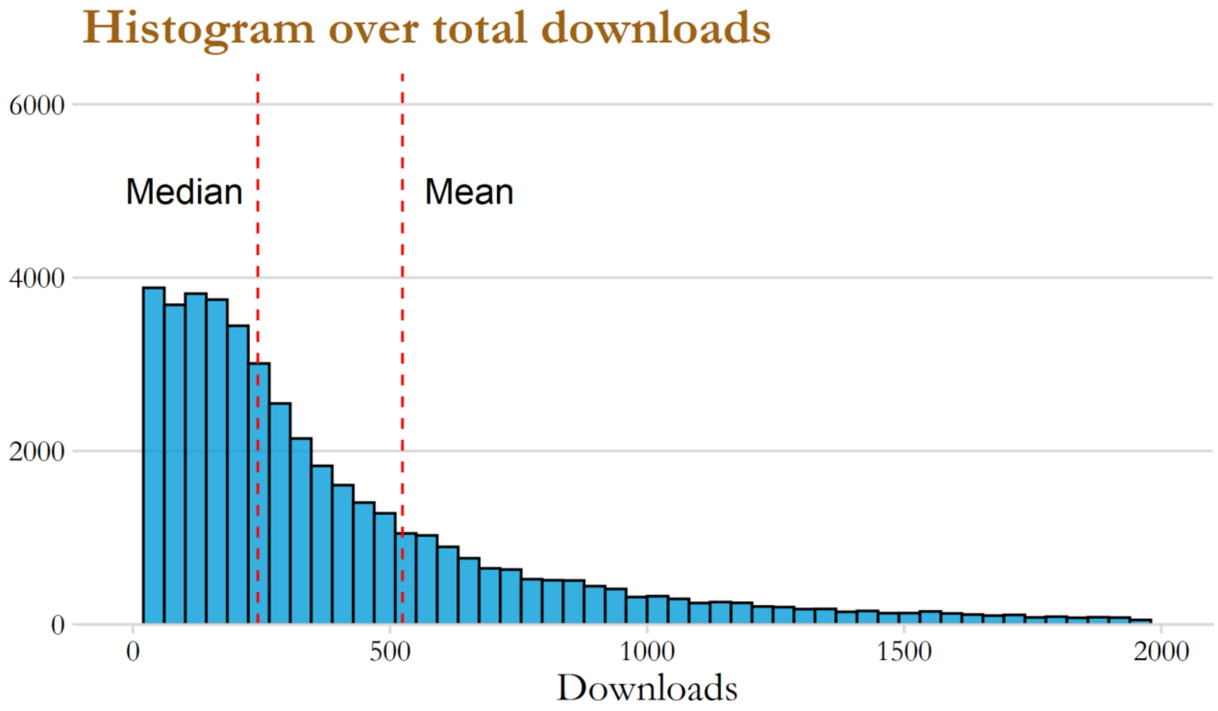


Figure 14: Histogram over total downloads

The first thing to notice is the large difference between the median and the mean. The mean is more affected by outliners in the data (James et al, 2013), which explains why the mean is around 500 downloads while the median is around 250 downloads.

2.7.1. Faculties and the average download

Let us compare if there seems to be a difference in how many times a paper on average is downloaded across the six faculties.

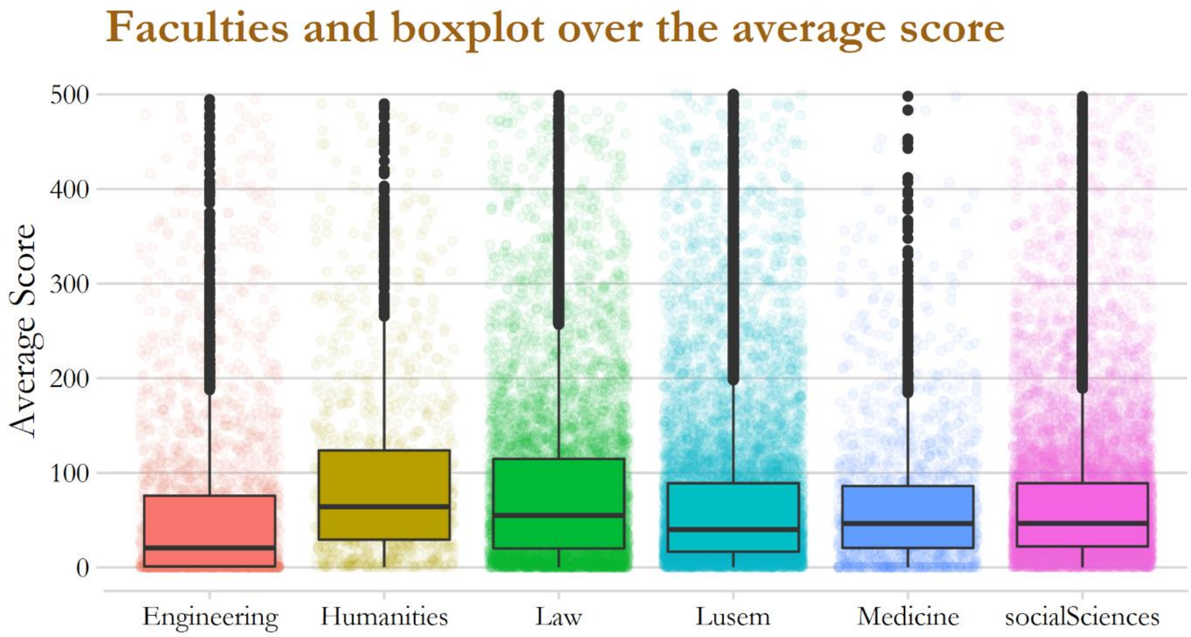


Figure 15: Boxplot over Faculties and their average downloads

The boxplot shows that the median for all the faculties regarding average download is almost the same. Humanities is the faculty with the highest median download while Engineering is the lowest. One of the reasons why Humanities show the highest median score might be because the faculty also produce the least amount of student papers, which leads to less papers with very few downloads that lower the median.

3. Classifying Faculties based on abstract

This part of the thesis will focus on text classification in order to classify which faculty an abstract belongs to. According to Ledolter (2013) so is the process of selecting and assembling the data that will be used in statistical models the most important step. The model will not produce good results if the underlaying data is bad or is not representative.

The abstracts that can be used are only those that are written in English because the packages in R can only handle English words when cleaning the data. Table 2 below demonstrates that there is a large difference in how many English papers each faculty have published.

Table 2: Total number of papers in English

Faculty	Published papers in English
LUSEM	4 347
Social Sciences	3 331
Engineering	1 806
Law	587
Humanities	291
Medicine	109

The three faculties Law, Humanities and Medicine have very few observations compared to the other faculties and can therefore not be used since the data will not be enough to create strong indications for classification (James et al, 2013; Ledolter, 2013). The faculties that will be used in this analysis are therefore LUSEM, Social Sciences and Engineering since they are the only ones with enough data when the constrain for English is applied.

Not all of the data that is available will be used, 1 500 abstracts will be randomly selected from each of the three faculties so that every faculty have the same number of abstracts so that none will be overrepresented in the dataset.

Table 3: Number of abstracts used in the model

Faculty	Number of observations
LUSEM	1 500
Social Sciences	1 500
Engineering	1 500

3.1. Cleaning the text data

This subsection will go through how the text data is preprocessed in order to create a classification model.

It is of importance to clean the text data by removing stop-words and reduce the words to its root-form since these procedures often increases accuracy for text classification models (Saif, Fernández, He, & Alani, 2014). Removing stop-words means that words such as *"where", "when", "to"* and *"at"* are removed from the abstracts. Another positive thing regarding removing stop-words is that the words does not contain much information that is helpful when classifying and the removal makes the model faster since there are less words.

R is case sensitive, so all the text was transformed to lowercase so that there will not be a difference between a word that is spelled with lowercase or uppercase letters.

The next step was to transform the words in the abstracts into their root-form, this means that words like *"looking"* will have the suffix *"ing"* removed from the word so all that is left is the root-form which in this case is *"look"*. The reason for using stemming is to transform inflectional variation of the word into the root word version, this makes it easier for the model since variations of the words becomes the same (Saif et al, 2014; Ledolter, 2013).

The following step was to extract tokenization of every word and create a document-term matrix. Tokenization means that every word is extracted from the abstracts in order to create a database of every word. These words are then used in a document-term matrix which is a matrix where the rows correspond to the documents (in this case the abstract) and the columns corresponds to words. The table below shows an example the document-term matrix:

Table 4: Example of the document-term matrix

	calcul	**chain**	**compani**	**cost**	**countri**
Abstract 1	0	0	1	0	0
Abstract 2	1	1	0	0	1
Abstract 3	0	0	0	0	0

The first document-term matrix had 14 827 words which was a very big matrix with a lot of words that does not contain much information. An algorithm was therefore used in order to remove the words that only occurred a few times. The number of words in the new document-term matrix was 3 619, this is the one that will be used in the classification model.

It is good practice to first investigate if the collected data seems to show signs that there is a difference between the faculties that makes it possible to classify them before starting to create a model. One approach is to study the most used words in each class, see table 5 below which displays the most common words for the faculties (not in their root-form).

Table 5: Summary over the ten most common words for each faculty (pre stemming)

Engineering		**LUSEM**		**Social Sciences**	
Word	**n**	**Word**	**n**	**Word**	**n**
thesis	1 001	study	939	study	1 280
model	617	thesis	685	thesis	709
system	582	market	677	social	650
purpose	558	purpose	627	analysis	460
control	542	model	513	eu	452
study	503	data	469	theory	440
process	462	economic	453	political	410
project	408	research	448	purpose	387
development	388	countries	420	development	378
design	385	paper	395	policy	274

Many of the most common words in table 5 are the same across the faculties, words like "*study*", "*thesis*", "*purpose*" and "*model*". This is a problem because the classification model wants words to be associated with a specific faculty, so it is easier to classify the abstract based on those words.

A better method to see words that are important for the corresponding faculty is instead to use term frequency–inverse document frequency (TF-IDF). The TF-IDF is calculated with the below equation where *"t"* is the word in the document *"d"* from the document set *"D":*

$$tfidf(t, d, D) = tf(t, d) * idf(t, D) \tag{1}$$

Where:

$$tf(t, d) = \log(1 + freq(tnd)) \tag{2}$$

$$idf(t, D) = log\left(\frac{N}{count(d \in D: t \in c)}\right) \tag{3}$$

The TF-IDF is often better to look at because it produces an improved understanding of the more important words that are likely to provide good information for classification models when predicting between classes (Bafna, Pramod, & Vaidya, 2016). The 15 words with the highest TF-IDF for the faculties can be viewed in figure 16 below:

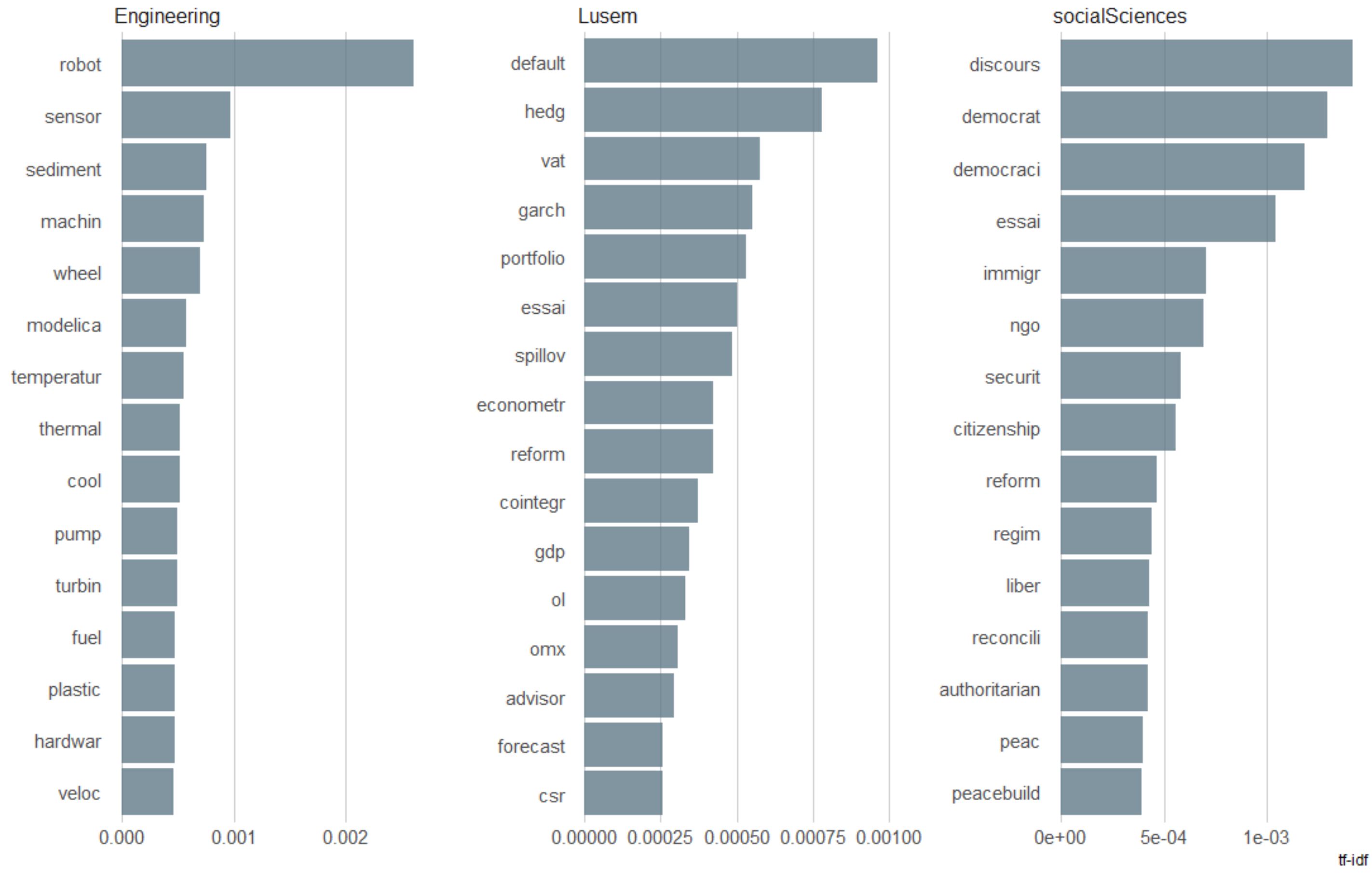


Figure 16: TF-IDF diagram over 15 words for each faculty

Comparing figure 16 with table 5 makes it clear that TF-IDF is better at showcasing interesting words, all the words are unique and has a clear connection to the corresponding faculty. A good example is the words "*default*", "*hedg*" and "*portfolio*" which are terms often used in economics and is therefore good representation for the faculty LUSEM.

3.2 Random Forest as the Classification Model

The focus of this thesis is not to perform or experiment with creating the optimal classification model. The emphasis is to see if a classification model creates good accuracy which would therefore suggest that the abstracts between the three faculties have notable differences.

Random Forest (RF) is the classification algorithm that will be used in this thesis. RF is a classic and powerful supervised method that can be used for regression and classification (James et al, 2013, Ledolter, 2013). RF have an advantage over regular decision trees because RF is an ensemble process that avoid over-fitting by merging several decision trees in a stochastic manner. Breiman (2001) argues that by using several independent that vote for classification creates better accuracy and less over-fitting when studying the majority vote compared to using a single decision tree.

Another advantage with RF is that it often performs well with large data that contains many features. However, one clear disadvantage is that RF is a black-box method, meaning that it is difficult to interpret how the model is thinking and what the classification is based on (James et al, 2013).

Other popular models for text classification such as Naïve Bayes, Support Vector Machines or Neural Networks will not be tested in this thesis due to time constraint and because difference in performance is not something this thesis intends to evaluate.

4. Empirical Findings and Analysis

This section presents the results from the RF-model created in the previous chapter. The results are then analyzed by using a confusion matrix and evaluating the model with metrices such as sensitivity, specificity and balanced accuracy.

4.1. Variable importance

RF is considered a black-box method, one of the only interpretations to understand how the model works is by investigating the variable importance which measures the most important variables for the model.

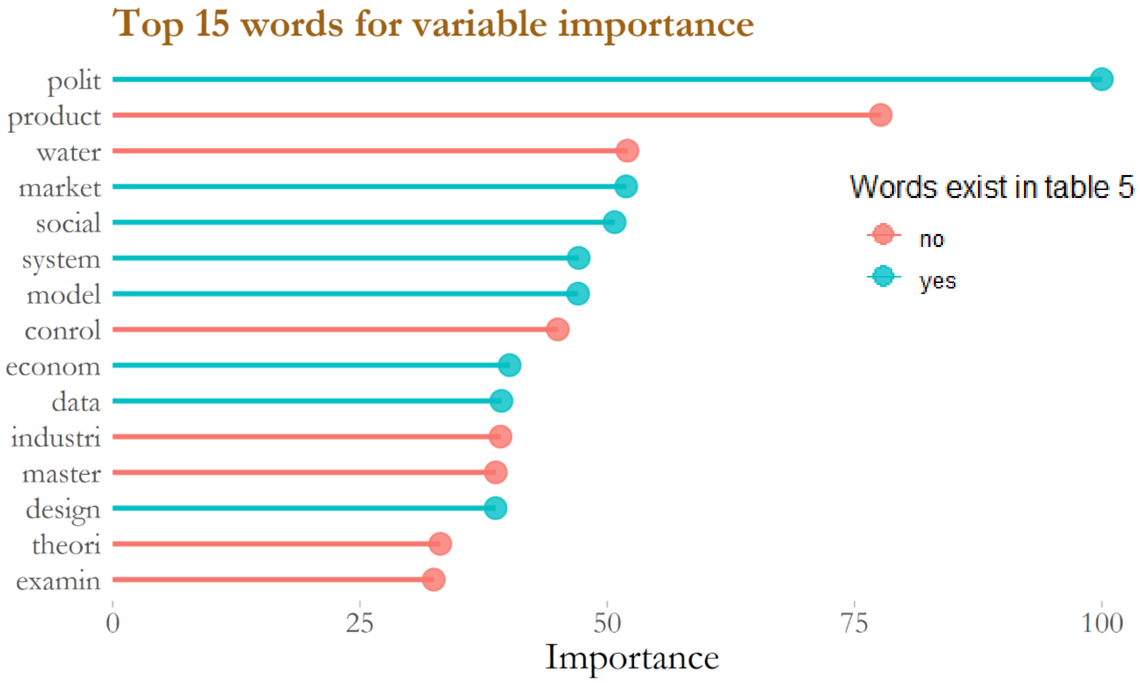


Figure 17: Variable importance from the RF-model

The three most important words for classifying the faculties are according to the above figure "*polit*", "*product*" and "*water*". It is interesting to see that eight out of the 15 words exist in table 5 but no words exist in figure 16 concerning the TF-IDF diagram. This would suggest that common words are important for the classification model.

4.2. Confusion matrix and statistics over model performance

The RF-model created in section 3.2 is used in order to predict the faculties for the testdata. If the model is perfect it would predict 900 correct predictions, 300 predictions for each faculty. The results are summarized in a confusion matrix in table 5 below. A confusion matrix is a simple way to create a clear understanding of the prediction across the groups and see if there is some group that performs worse or better (Ledolter, 2013).

Table 6: Confusion matrix for the classification model; Accuracy = 0.8022

		Reference	
	Engineering	**LUSEM**	**Social Sciences**
Predicted Engineering	260	44	18
Predicted LUSEM	31	219	39
Predicted Social Sciences	9	37	243

The accuracy for the model is 0.8022 which is quite good considering that no special care has been taken in order to optimize the model.

By observing the confusion matrix, we can see that the model rarely predicts Social Sciences for Engineering (happened 9 times) or the opposite (happened 18 times). This would suggest that Social Sciences and Engineering are the two faculties that are most unlike since they rarely get classified as the other faculty.

LUSEM is the faculty that has the worst accuracy, only 219 out of 300 observations was correctly identified. The problem with LUSEM seems to be that it has a high misclassification with the two other faculties, especially Engineering.

4.3. Summary statistics over k-fold cross-validation

The R-package Caret was used when creating the RF-model and Caret makes it able to easily perform k-fold cross-validation when training the model. Cross-validation means that a small subsample of the data is used as testdata and predicted on using the rest of the traindata. The K split the data into equal big parts as the number K (James et al, 2013, Ledolter, 2013). See figure 18 below that illustrates how k-fold cross-validation works for k = 5.

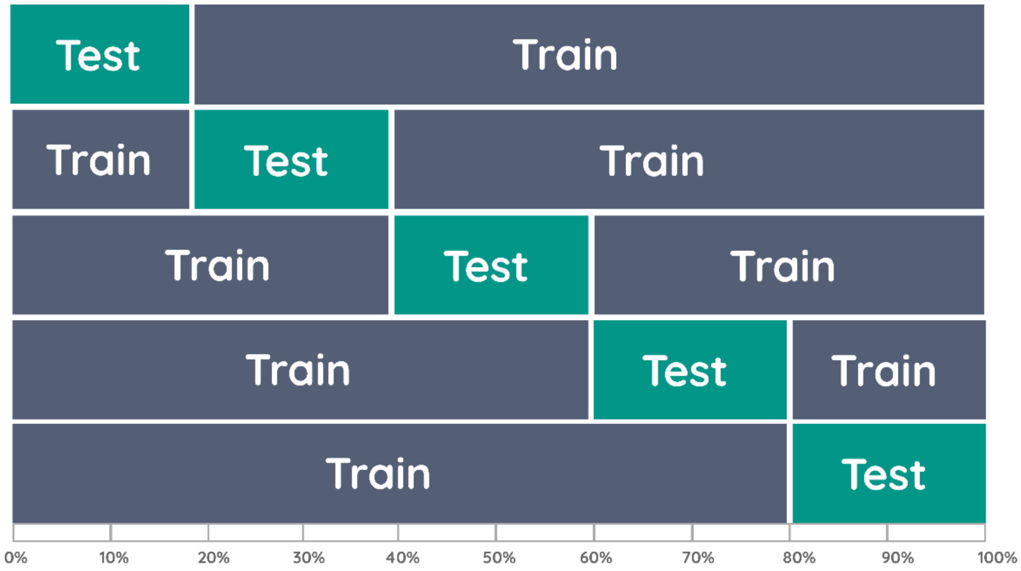


Figure 18: Example of k-fold cross-validation where k = 5

The good thing about using k-fold cross-validation is that Caret provides summary statistics that evaluate how the model have performed under these predictions and thus give good estimates on how the model generally performs when predicting on new data. Table 7 below shows summary statistics over performance metrices such as sensitivity, specificity and balanced accuracy that was calculated under the k-fold cross-validation.

Table 7: Summary over model statistics for the faculties

Statistics	**Engineering**	**LUSEM**	**Social Sciences**
Sensitivity	0.8667	0.7287	0.8113
Specificity	0.8980	0.8827	0.9227
Balanced Accuracy	0.8823	0.8057	0.8670

Sensitivity is a measurement to show how good the model is at predicting the corresponding class it is calculated by using the below equation:

$$Sensitivity = \frac{Number\ of\ true\ positives}{number\ of\ true\ positives + number\ of\ false\ negatives} \tag{4}$$

The faculty with the lowest sensitivity is LUSEM which is in line with the findings in the confusion matrix (see table 6).

Specificity is a measurement used to determine how good the model is at classifying the actual negatives for the corresponding class, it is computed using the below formula:

$$Specificity = \frac{Number\ of\ true\ negatives}{number\ of\ true\ negatives + number\ of\ false\ positives} \qquad (5)$$

The specificity between the faculties are closer to each other compared to the spread in sensitivity. The highest specificity has Social Sciences and that becomes clear when observing table 6 which show that there are not many predictions for Social Sciences for the other faculties.

The balanced accuracy in table 7 shows that LUSEM have the lowest accuracy which agrees with table 6. The accuracy for Engineering is the highest and is 0.8823 and Social Sciences is little behind with an accuracy of 0.8670.

5. Conclusion

This section will recapitulate the thesis and examine the results and insights from the visualizations with consideration to the research purposes that was stipulated in the first chapter. The chapter will end with suggestion that can be used for further research.

5.1. Summary

The purpose of this thesis was to examine and create insights concerning student papers published from Lund University at the website LUP. The dataset was created by the author of this thesis using web-scraping in R. The second purpose of the thesis was to try and see if it was possible to use the abstract from the student papers in order to classify them to their faculty.

The main emphasis of chapter 2 was to use visualizations in order to fulfill the first purpose. The thesis showcased that the number of publications has steadily increased for every year except for the last three where the publications stays around 4 200 per year. Figure 3 shows that bachelor and two-year master is the most published papers for each year and one-year master seems to get less publications over time.

English has become the most common written language per year and is most popular in the master papers (see figure 5 and 6). The most common keywords such as "*management*", "*social*", "*law*" and "*företagsledning*" reflects that the most published faculties are Social Sciences, LUSEM and Law (see figure 7 and 9).

Figure 14 that shows the total download of the papers shows that the distribution is very skewed. There are only a few percent of the papers that gain a large number of downloads, this might be because most of the papers are in Swedish and student papers are mostly read by other students and not by the broad public (Newkirk, 1984).

A random forest model was created in order to classify the abstract to the three faculties LUSEM, Social Sciences and Engineering. The confusion matrix on the testdata got an accuracy of 0.8022 but showcased that LUSEM was the hardest faculty to classify. The k-fold cross-validation results also agreed on that LUSEM was the faculty with the lowest classification accuracy. One reason for why LUSEM might be the hardest to classify is that topics within economic/business are very broad and can be close to subjects that occur in social sciences or engineering.

5.2. Further research

The focus of this thesis was not to create an optimal classification model, it was simply to go through the approach of cleaning the data and test if a model could classify the abstracts with a moderately high accuracy.

Further research could therefore use the same dataset and instead focus on using different models and optimizations in order to investigate how the models performs and which model that yields the best result.

References:

Bafna, P., Pramod, D., & Vaidya, A. (2016). Document clustering: TF-IDF approach. *International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT),* 61-66.

Björkman, B. (2008). 'So where we are?' Spoken lingua franca English at a technical university in Sweden. *English Today*, 35-41.

Breiman, L. (2001). Random forests. *Machine Learning, 45(1)*, 5-32.

Cinamon, R. G., & Rich, Y. (2002). Gender Differences in the Importance of Work and Family Roles: Implications for Work–Family Conflict. *Sex Roles*, 531-541.

Healy, K. J. (2019). *Data visualization : a practical introduction.* Princeton, New Jersey: Oxfordshire.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning : with applications in R.* New York: Springer.

Ledolter, J. (2013). *Data mining and business analytics with R.* Hoboken, NJ: Wiley.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* London: The MIT Press.

Newkirk, T. (1984). How students read student papers: An exploratory study. *Written Communication 1.3*, 283-305.

Saif, H., Fernández, M., He, Y., & Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of Twitter. *Ninth International Conference on Language Resources and Evaluation*, 810-817.

SCB. (2020, Augusti 4). *statistiska centralbyrån*. Retrieved from SCB: https://www.scb.se/hitta-statistik/statistik-efter-amne/utbildning-och-forskning/hogskolevasende/studenter-och-examina-i-hogskoleutbildning-pa-grundniva-och-avancerad-niva/

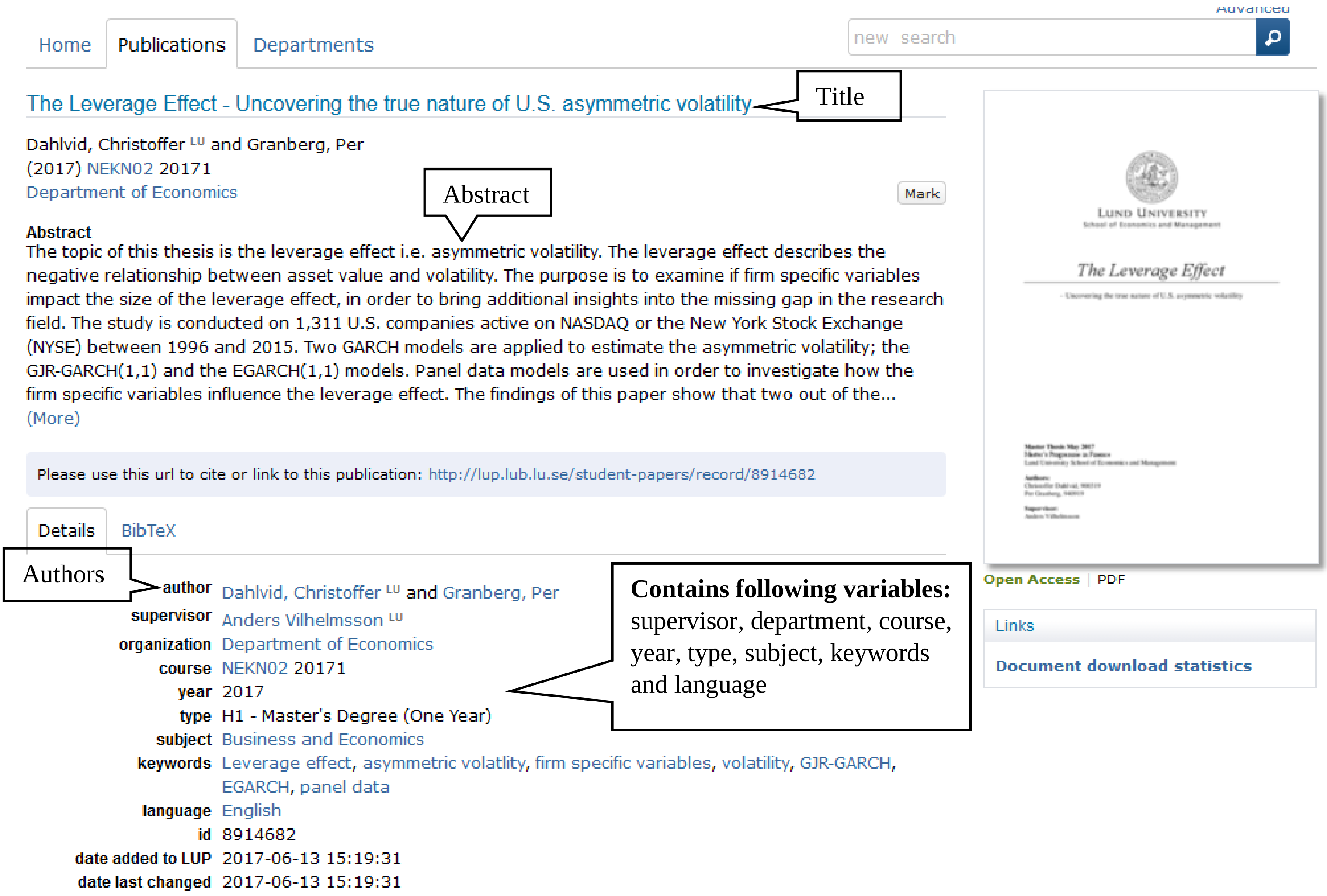
Schilt, K., & Wiswall, M. (2008). Before and After: Gender Transitions, Human Capital, and Workplace Experiences . *The B.E. Journal of Economic Analysis & Policy, 8(1).*

Vargiu, E., & Urru, M. (2013). Exploiting web scraping in a collaborative filtering based approach to web advertising. *Artificial Intelligence Research*, 44-54.

Zaki, M. j., & Meira, W. (2014). *Data Mining and Analysis Fundamental Concepts and Algorithms.*

Appendix A – LUP website information

This appendix will demonstrate how the information from LUP looks like and some of the variables that was collected. The first picture below shows information regarding the thesis, the picture on the next page will instead show statistics regarding how many times the thesis has been downloaded.


Advanced
Home
Publications
Departments
new search
Mark
LUND UNIVERSITY
School of Economics and Management
The Leverage Effect
- Uncovering the true nature of U.S. asymmetric volatility
Open Access | PDF
Links
Document download statistics

Title

Abstract

Authors

Contains following variables: supervisor, department, course, year, type, subject, keywords and language

The Leverage Effect - Uncovering the true nature of U.S. asymmetric volatility

Dahlvid, Christoffer LU and Granberg, Per
(2017) NEKN02 20171
Department of Economics

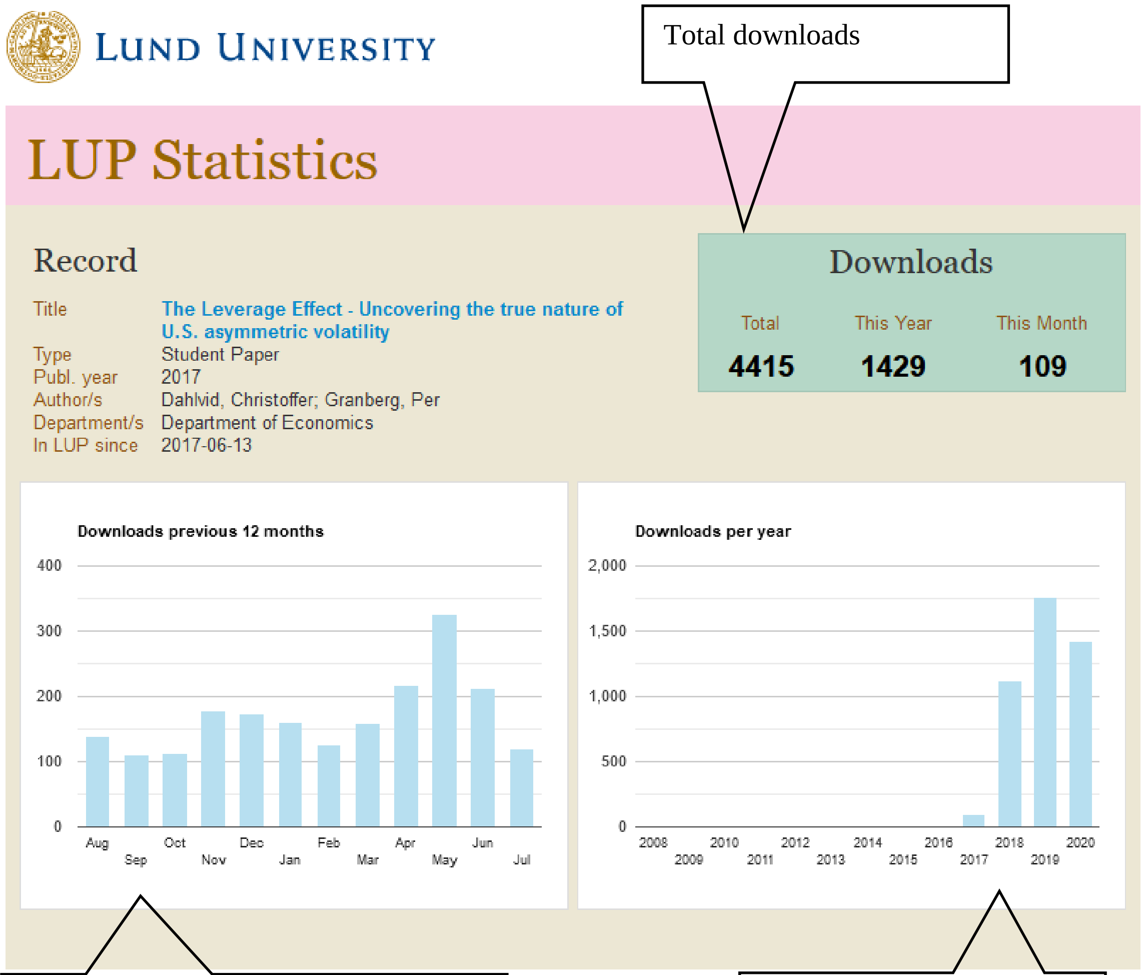
Abstract
The topic of this thesis is the leverage effect i.e. asymmetric volatility. The leverage effect describes the negative relationship between asset value and volatility. The purpose is to examine if firm specific variables impact the size of the leverage effect, in order to bring additional insights into the missing gap in the research field. The study is conducted on 1,311 U.S. companies active on NASDAQ or the New York Stock Exchange (NYSE) between 1996 and 2015. Two GARCH models are applied to estimate the asymmetric volatility; the GJR-GARCH(1,1) and the EGARCH(1,1) models. Panel data models are used in order to investigate how the firm specific variables influence the leverage effect. The findings of this paper show that two out of the...
(More)

Please use this url to cite or link to this publication: http://lup.lub.lu.se/student-papers/record/8914682

Details BibTeX

author	Dahlvid, Christoffer LU and Granberg, Per
supervisor	Anders Vilhelmsson LU
organization	Department of Economics
course	NEKN02 20171
year	2017
type	H1 - Master's Degree (One Year)
subject	Business and Economics
keywords	Leverage effect, asymmetric volatlity, firm specific variables, volatility, GJR-GARCH, EGARCH, panel data
language	English
id	8914682
date added to LUP	2017-06-13 15:19:31
date last changed	2017-06-13 15:19:31

Total downloads



LUND UNIVERSITY

LUP Statistics

Record

Title	The Leverage Effect - Uncovering the true nature of U.S. asymmetric volatility
Type	Student Paper
Publ. year	2017
Author/s	Dahlvid, Christoffer; Granberg, Per
Department/s	Department of Economics
In LUP since	2017-06-13

Downloads

Total	This Year	This Month
4415	1429	109

Downloads for the previous 12 months

Downloads for each year

Appendix B – Packages used in R

In this appendix I will describe some of the packages I used in this thesis.

Packages used for visualization and cleaning data:

ggplot2 (create most of the graphs)

cowplot (themes)

dplyr (clean/handle data)

naniar (creating figure 1 regarding missing values)

scales (for handling scales, example is figure 4)

viridis (for different color in the plots)

rvest (used to scrape the data form LUP)

Packages used for text classification:

caret (random forest and k-fold classification)

tidytext (handle text data)

SnowballC (stemming the words)

Stopwords (removing stop words)

Appendix C – 15 Random Names and gender

This appendix shows 15 random selected first names from the dataset and the corresponding gender according to the r-package Genderize.

Name	Gender according to Genderize
Anton	Male
Dennis	Male
Marie	Female
Olivia	Female
Theodor	Male
Ellen	Female
Yvonne	Female
Carolina	Female
Zhou	Missing value
Linus	Male
Victoria	Female
Ingegerd	Missing value
Camilla	Female
David	Male
Josefine	Female

The table above show that there are two missing values, this happens when Genderize does not know what gender to apply. After looking at other missing values concerning genders, I noticed that this often happens to old Swedish names or names from other nationalities such as China.