Covid-19 Pandemic and the World of Finance

Rolf Beutner, Bamberg, Bavaria, Germany

2021-07-10

Contents

1 Overview

This report is part of the HarvardX course: **PH125.9x Data Science** and contains the results of the final capstone project.

It consists of **five parts**, starting with this **Overview**, followed by a brief **Introduction and Motivation** outlining the research questions.

The **Summary** section describes how to download, prepare and cleanse the required data sets for Covid-19-, financial- and country-data from:

- "International Monetary Fund (IMF)" (2021)
- "Our World In Data (OWID)" (2021)
- "Software Repository Accounting and Finance (SRAF), University of Notre Dame (IN,USA)" (2021)
- "Tidyquant - Bringing Financial and Business Analysis to the Tidyverse, Daily Stock Prices (and Many Other) from Federal Reserve Economic Data (FRED)" (2021)
- "Worldbank Institute (WBI)" (2021)

The next section **Investigation** describes the exploratory data analysis performed to get an overview and initial approaches to the given data.

In the **Methods and Analysis** section the data sets were taken from the investigation section, combined and a sentiment analysis was done. The data sets were then enriched by countries region, sub-region and income-level to order-, group- and filter. With this combined data set we investigate deeply the detail texts which were sentiment analyzed before and use different machine learning algorithm to verify whether these algorithms can predict the sentiments too.

The **Results** section orders, groups, and filters the output of the methods and analysis section to display it from different perspectives.

The report ends with a **Conclusion** on the results and possible next steps followed by appendix with abbreviations and references.

Hints:
The code is included in the RMD file but explicitly switched off (echo=FALSE).
My Environment:

- Windows 10
- R-3.6.3-win
- Rtools35
- RStudio-1.4.1106
- Ghostscript gs9540w64

---

2 Introduction and Motivation

When searching for a topic for my capstone project, I became interested in a connection between the current **Covid-19 pandemic** and the **financial world**. When you see the performance of their most important index, the "Dow Jones Industrial Average" (**DOW** or DJIA) one year after the start of the global pandemic in February 2020 this is stronger than ever - see Figure 1. The financial world seems to be the winner of the crisis.

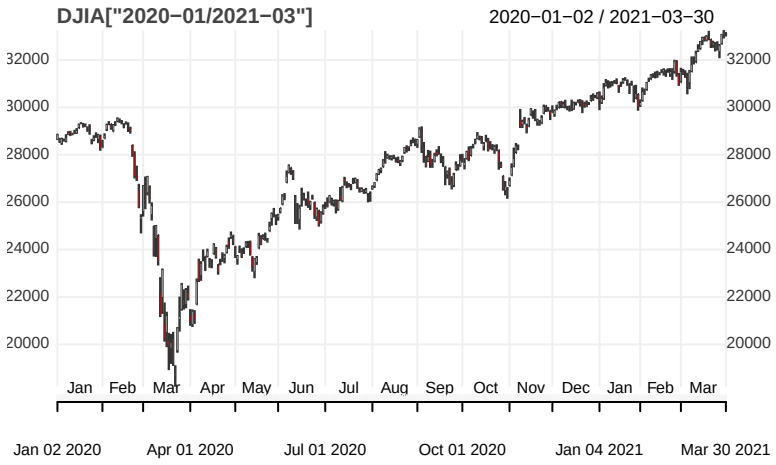


Figure 1: Dow Jones Index Jan 2020 - Mar 2021

Research Questions:

Focussing on the financial world, what were and are the actions and policy measures related to the Covid 19 pandemic?

Were these actions and policy measures positive or negative or have other sentiments?
What is in this domain "positive" ?
Can machine learning be used to evaluate the sentiments of financial texts?
How do the machine learning algorithms differ in quality and speed?

Does the country's region in the world have an impact?
Do different income levels lead to different measures?
What are common measures regardless of region or income level?

A review by the Word Bank Institute's Center for Finance, Competitiveness & Innovation [FCI] helped to get into it. ("The World Bank Group's Response to the COVID-19 (Coronavirus) Pandemic" 2021)

2.1 Summary and Overview of the data

Data sets
The data sets, compiled by the International Monetary Fund (IMF), OurWorldInData.org (OWID) and World Bank Institute (WBI) were merged, grouped, aggregated and sentiment-analyzed based on an financial analysis classification set,[1] enriched and filtered by region, sub-region and income-level - see Figure 2. Based on that sentiment analysis different machine learning algorithms were used to find the best and fastest prediction.

In the following a data flow diagram, generated by graphviz / DiagrammR:[2] [3]

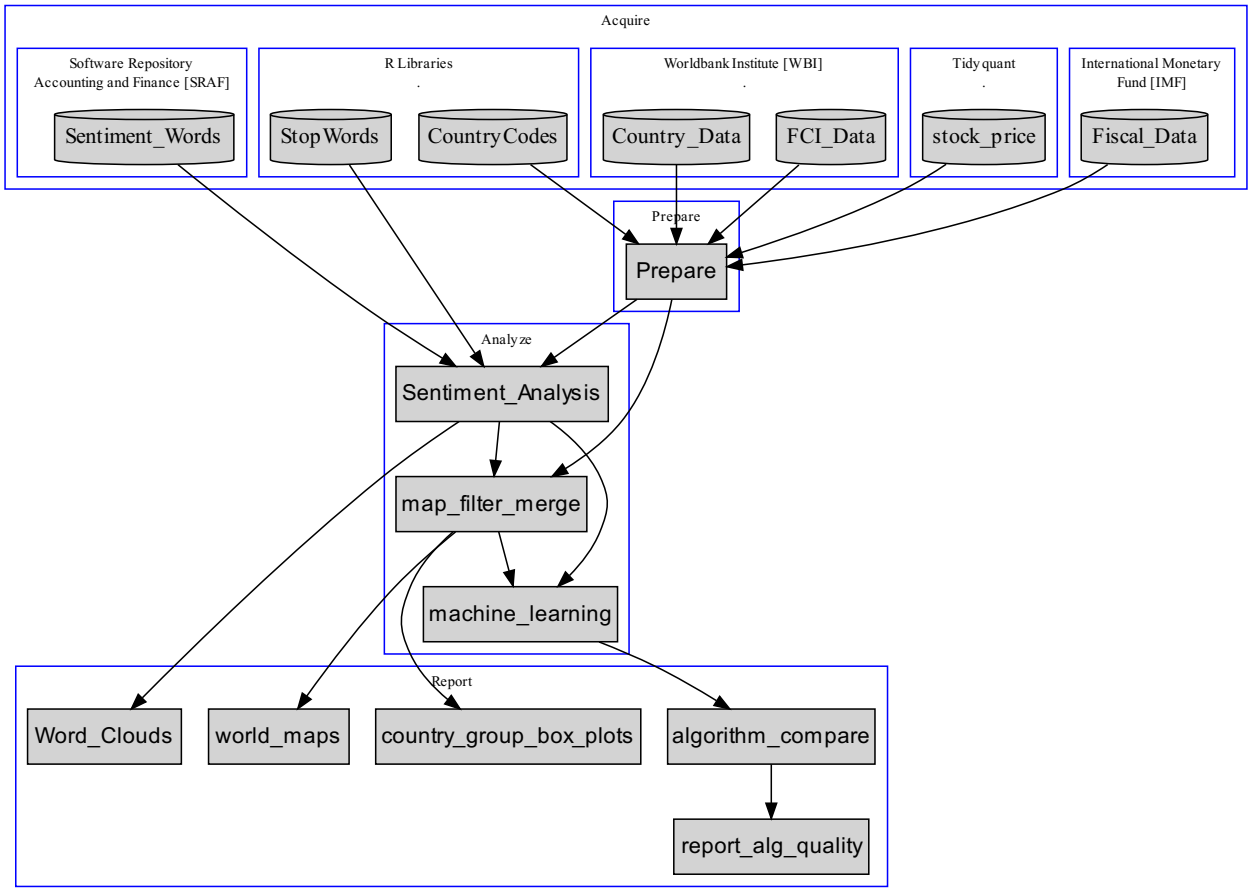


Figure 2: Overview of Data Input and Flow

A. **Financial data**

1. **FCI data** by **Word Bank Institute** [WBI]

   Finance, Competitiveness & Innovation [FCI] Global Practice overview of policy measures
   Data taken in jurisdictions and by type of measure in support of the financial sector to address the impact of the COVID-19 pandemic:
   ("COVID-19 Finance Sector Related Policy Responses Catalog" 2021)
   ("COVID-19 Finance Sector Related Policy Responses Data Link / Covid-Fci-Data.xlsx" 2021)
   3672 rows x 14 columns, 148.000 detail words

[1] "Software Repository Accounting and Finance / LM Sentiment Word Lists, University of Notre Dame (IN,USA)" (2018)
[2] https://dreampuf.github.io/GraphvizOnline
[3] https://graphviz.org/doc/info/shapes.html

2. **Fiscal Measure Response data** by **International Monetary Fund** [IMF]

   Additional country data - expenses for Covid-19 of GDP in percent.
   ("Fiscal Policies Database / Fiscal-Policies-Database-in-Response-to-COVID-19, International Monetary Fund (IMF)" 2021)
   193 rows x 25 columns

3. **DOW or DJIA** by **Yahoo finance [YF]**

   Dow jones industrial average" (DOW or DJIA) index time series value as motivating example above by accessing yahoo finance via R tidyquant package. The data is only used in the motivation section above.
   313 rows x 6 columns

B. **Country and Locations data**

1. **Country and Lending Group** by **Word Bank Institute** [WBI]

   Additional country data - grouping and ratings
   ("Country and Lenting Group of WBI - Knowledge Base" 2021)
   ("World Bank List of Economies (June 2020), Country and Lenting / CLASS.xls" 2021)
   279 rows x 9 columns

C. **Sentiment Resources**

1. **Textual Analysis Resources**, Software Repository Accounting and Finance [SRAF] by ("Software Repository Accounting and Finance (SRAF), University of Notre Dame (IN,USA)" 2021)

2. **Sentiment Words Financial** ("Software Repository Accounting and Finance / Textual Analysis / Resources, University of Notre Dame (IN,USA)" 2018)
   4150 words + sentiment

Preparations: The mentioned data sets from the different locations in Excel- and CSV-format were downloaded and transferred into data frames by using the tidy packages tidyverse, tidytext, tidyquant,

- cleaned empty/white/marked as empty values especially in the excel sheets,
- removed not needed features,
- renamed/normalized feature names and values to make data sets comparable/mergeable/readable,
- transposed data formats e.g. dates to year/month pairs, absolute to relative values, exchange of row/column, use pivot tables,
- prepared filters and mappings.

Problems: Complex formats in Excel like tables within tables (Worldbank- and IMF-Excel), hidden tabs, many features look interesting but lead beyond the scope and into complexity.
The decision what to delete or aggregate to not lose information or get misleading results was a challenge.
Merging data across table by free-text strings required precise preparation before it worked - different notation for the same country. A big helper was the R library "countrycodes" to unify country identification.
Sometimes the latex generation messages and problems were annoying, seem non-printable character caused the problems. Unclear, why some fig.cap below graphs do not occur. And the "Label multiply defined" warning is a mystery.

2.1.1 Financial Resources

2.1.1.1 FCI (= Finance, Competitiveness & Innovation) data related to Covid-19 by World Bank Institute [WBI]

This data set covers the time of February 2020 - March 2021.[4] ("COVID-19 Finance Sector Related Policy Responses Data Link / Covid-Fci-Data.xlsx" 2021)

```
## # A tibble: 6 x 14
##      ID `Country Name`     `Country ISO3` `Income Level`      Authority
##   <dbl> <chr>              <chr>          <chr>               <chr>
## 1     1 China              CHN            Upper middle income SUP
## 2     2 Canada             CAN            High income         CB
## 3     3 China              CHN            Upper middle income CB
## 4     4 Thailand           THA            Upper middle income CB
## 5     5 Russian Federation RUS            Upper middle income CB
## 6     6 Belarus            BLR            Upper middle income CB
## # ... with 9 more variables: Date <dttm>, Level 1 policy measures <chr>,
## #   Level 2 policy measures <chr>, Level 3 policy measures <chr>,
## #   Details of the measure <chr>, Reference <chr>, Termination Date <dttm>,
## #   Modification of Parent Measure <chr>, Parent Measure <dbl>
```

```
## [1] "FCI-Data: 3884 rows x 14 columns"
```

2.1.1.2 Fiscal measure response data related to Covid-19 by International Monetary Fund [IMF]

("International Monetary Fund (IMF) COVID-FM-Database" 2021)

This database summarizes key fiscal measures governments have announced or taken in selected economies in response to the COVID-19 pandemic as of March 17, 2021. It includes COVID-19 related measures since January 2020 and covers measures for implementation in 2020, 2021, and beyond. We are using the key features "Budget in % of GDP" and "Budget in USD"

> Data are used from **January 2020 to March 2021** because the FCI-data set covers only this time period.

```
## Warning in fansi::strwrap_ctl(x, width = max(width, 0), indent = indent, :
## Encountered a C0 control character, see `?unhandled_ctl`; you can use
## `warn=FALSE` to turn off these warnings.
```

```
## # A tibble: 6 x 25
##   `G20 + Spain` `Country Group` `Country /1` `Government Level` ...5  Unit...6
##   <chr>         <chr>           <chr>        <chr>               <lgl> <chr>
## 1 1             AE              Australia    General Government NA    LC bn
## 2 <NA>          <NA>            <NA>         <NA>                NA    USD bn
## 3 <NA>          <NA>            <NA>         <NA>                NA    % GDP
## 4 1             AE              Canada       Central Government NA    LC bn
## 5 <NA>          <NA>            <NA>         <NA>                NA    USD bn
## 6 <NA>          <NA>            <NA>         <NA>                NA    % GDP
## # ... with 19 more variables: Total on-budget  (A-D) <dbl>, Total size <chr>,
## #   Additional spending and forgone revenue in the health sector <chr>, Total
## #   size...10 <chr>, Additional spending and forgone revenue
```

[4] "COVID-19 (Coronavirus) Response" (2021)

```
## #   in areas other than health <chr>, Total  size...12 <chr>,
## #   D. Accelerated spending and deferred revenue in areas other than health <chr>,
## #   ...14 <lgl>, Unit...15 <chr>, Total off-budget (B+C) <chr>, Total
## #   size...17 <chr>,
## #   Equity injections, asset purchases, loans, debt assumptions, including through extra-budgetary fu
## #   ...19 <lgl>, Unit...20 <chr>, Total  size...21 <chr>,
## #   Guarantees (on loans, deposits etc.) <chr>, Total  size...23 <chr>,
## #   Quasi-fiscal operations (noncommercial activity of public corporations on behalf of government) <
## #   Automatic stabilizers, contingency lines <chr>
```

```
## [1] "Fiscal: 193 rows x 25 columns"
```

2.1.2 Country + Locations Resources

2.1.2.1 Country Data by Word Bank Institute [WBI] The region, sub-region and income-level are included in this data. The income-level is not well filled, only 155 of 251 rows. The country class data set is better filled for income-level, so we enrich / coalesce the both features. ("Countries with Regional Codes / All.csv" 2021)

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   name = col_character(),
##   `alpha-2` = col_character(),
##   `alpha-3` = col_character(),
##   `country-code` = col_character(),
##   `iso_3166-2` = col_character(),
##   region = col_character(),
##   `sub-region` = col_character(),
##   `intermediate-region` = col_character(),
##   `region-code` = col_character(),
##   `sub-region-code` = col_character(),
##   `intermediate-region-code` = col_character()
## )
```

```
## [1] "Country: 249 rows x 11 columns"
```

2.1.2.2 Country Class Data by Word Bank Institute [WBI] Here the countries are classified additionally in financial regions. The financial- or income-level, here named income-group, is better filled as in the previous data. ("World Bank List of Economies (June 2020), Country and Lenting / CLASS.xls" 2021)

```
## # A tibble: 5 x 9
##   x...1 x...2 Economy  Code  X     Region  `Income group` `Lending catego~ Other
##   <chr> <chr> <chr>    <chr> <chr> <chr>   <chr>          <chr>            <chr>
## 1 47    <NA>  Costa R~ CRI   <NA>  Latin ~ Upper middle ~ IBRD             <NA>
## 2 1     <NA>  Afghani~ AFG   <NA>  South ~ Low income     IDA              HIPC
## 3 70    <NA>  French ~ PYF   <NA>  East A~ High income    ..               <NA>
## 4 107   <NA>  Kuwait   KWT   <NA>  Middle~ High income    ..               <NA>
## 5 50    <NA>  Cuba     CUB   <NA>  Latin ~ Upper middle ~ ..               <NA>
```

```
## [1] "Country Class: 275 rows x 9 columns"
```

2.1.3 Sentiment Analysis Resources

2.1.3.1 Financial Sentiment Words by Software Repository Accounting and Finance [SRAF]

The following resources were evaluated.
The builtin data in R was sufficient - no need to download, here for reference and as background.

("Software Repository Accounting and Finance (SRAF), University of Notre Dame (IN,USA)" 2021) ("Software Repository Accounting and Finance / Textual Analysis / Resources, University of Notre Dame (IN,USA)" 2018)

> "... almost **three-fourths of the words** identified as **negative by the widely used Harvard Dictionary** are words **typically not considered negative in financial contexts**. We develop an alternative negative word list, along with five other word lists, that better reflect tone in financial text. We link the word lists to 10-K filing returns, trading volume, return volatility, fraud, material weakness, and unexpected earnings"
> (T.Loughran, B.McDonald 2010)

Examples for these lists:[5][6]

- *negative*: restated, litigation, termination, discontinued, penalties, unpaid
- *positive*: efficient, improve, profitable, upturn
- *uncertain*: approximate, contingency, depend, fluctuate, indefinite, risk, uncertain, variability
- *litigious*: ("reflecting a propensity for legal contest"): regulated, legal, law
- *constraining*: claim, committing, limit, regulation, require

*This is the **fundamental basis** for further analysis **in this capstone project**.The detailed policy and measure texts of the countries are analyzed and categorized.The "Maximal Sentiment Category" (= category with the most words) determines the overall sentiment of the policies and measures for the country, the region, the sub-region or income-level.*

```
## # A tibble: 5 x 2
##   word        sentiment
##   <chr>       <chr>
## 1 deceased    negative
## 2 turbulence  negative
## 3 theretofore litigious
## 4 fictitious  negative
## 5 innovated   positive
```

```
## [1] "Sentiment_Classifier: 4150 rows x 2 columns"
```

[5] T.Loughran, B.McDonald (2010)
[6] "Documentation Loughran McDonald MasterDictionary (SRAF)" (2018)

2.2 First approaches to the data

In this chapter the data sets downloaded in the previous chapter are explored. That means they are shortly inspected, cleaned, features are selected or dropped, datas are firstly visualized - mainly as table.

2.2.1 FCI-data: prepare

The FCI data set is the main research table - the following preparation steps were done:

- renames, drop unused columns, cleanup whitespace(!!), inconsistent future dates
- rename cryptic and long content,
- normalize to capitalized measures

The data in this data set is **only defined for 156 countries** of the world. So there will be gaps in the sentiment-per-country-map.

```
## # A tibble: 5 x 10
##   Country  ISO3  IncomeLevel  Authority Date                Measure1  Measure2
##   <chr>    <chr> <chr>        <chr>     <dttm>              <chr>     <chr>
## 1 Germany  DEU   High Income  SUP       2020-12-15 00:00:00 Banking ~ Prudential
## 2 Sierra ~ SLE   Low Income   CB        2020-03-18 00:00:00 Liquidit~ Policy Ra~
## 3 Mexico   MEX   Upper Middl~ CB        2020-04-21 00:00:00 Liquidit~ Liquidity~
## 4 Malaysia MYS   Upper Middl~ CB        2020-02-27 00:00:00 Banking ~ Support B~
## 5 Philipp~ PHL   Lower Middl~ CB        2020-04-24 00:00:00 Liquidit~ Liquidity~
## # ... with 3 more variables: Measure3 <chr>, Detail <chr>, ym <chr>
```

```
## [1] "FCI data: 3821 rows x 10 columns"
```

2.2.2 Fiscal-data: prepare

The fiscal data table contains the monetary expenses a country has made as reaction to the pandemic, absolute in US-Dollar and relative to the countries' GDP. The following preparation steps were done:

- cleanup
- rename columns
- drop unused columns
- fill up empty cells caused by excel cell grouping - the most upper cell in that group has a value, can be filled down good by the fill function.

First lines of result table:

General Government = Central Government + State- + Local Government[7]

```
## # A tibble: 5 x 22
##   G20Spain Cgroup Country GovLevel  Unit1  Budget Tsize1  AddHealth      Tsize2
##   <chr>    <chr>  <chr>   <chr>     <chr>   <dbl> <chr>   <chr>          <chr>
## 1 0        AE     Belgium General ~ % GDP    8.03 1.8684~ "Additional s~ 6.1614~
## 2 0        LIDC   Kenya   Central ~ LC bn  250.   7.6     "Additional s~ 242
## 3 0        EM     Serbia  General ~ LC bn  308    73      "· 10 percent~ 235
## 4 0        EM     Chile   Central ~ USD bn  20.7  2.2154~ "Additional s~ 18.520~
```

[7] "Definitions of Government in IMF-Supported Programs" (2013)

```
## 5 0        AE     Belgium General ~ LC bn   36.1  8.4     "Additional s~ 27.7
## # ... with 13 more variables: AddNonHealth <chr>, Tsize3 <chr>,
## #   AccNonHealth <chr>, Unit2 <chr>, TBudgetBC <chr>, Tsize4 <chr>,
## #   EqInjAssPLD <chr>, Unit3 <chr>, Tsize5 <chr>, Guarantees <chr>,
## #   Tsize6 <chr>, LoanFundOther <chr>, QuasiFiscalOp <chr>
```

```
## [1] "Fiscal data: 189 rows x 22 columns"
```

2.2.3 Country data: prepare

The country table contains the information about ISO3-Code (important for unique merges to other datasets), region and sub-region. The country class table contains additionally a financial group, which we can map to the country via the ISO3 code.
The following preparation steps were done:

- cleanup
- rename columns
- drop unused columns
- fill up empty cells with the value of the cell above to handle excel grouped cells
- map FCI-data-income-level to the country - but has only 161 of 251 filled
- map / coalesce WBI-Class-data-Income-Group - ramp up to 218 filled income-levels
- replace / normalize the country names as used in other tables - to get it merged
- use only those rows where the region, sub-region and income-level is filled - otherwise it is not helpful for further analysis

Table 1: Country <-> Region/SubRegion/IncomeLevel

	ISO3	Country	Region	SubRegion	IncomeLevel	Economy	FinRegion	Incomegroup
1	ABW	Aruba	Americas	Latin America and the Caribbean	High	Aruba	Latin America & Caribbean	High
2	AFG	Afghanistan	Asia	Southern Asia	Low	Afghanistan	South Asia	Low
3	AGO	Angola	Africa	Sub-Saharan Africa	Lower Middle	Angola	Sub-Saharan Africa	Lower Middle
4	AIA	Anguilla	Americas	Latin America and the Caribbean	Upper Middle	NA	NA	NA
6	ALB	Albania	Europe	Southern Europe	Upper Middle	Albania	Europe & Central Asia	Upper Middle

```
## [1] "Country data: 218 rows x 8 columns"
```

2.2.4 World map: prepare

The ggplot world map is joined with the country data set. Unfortunately the world_map has no ISO3-code but we can enrich that to have a high quality feature. Further preparations:

- cleanup - some countries seem not not have a unique ISO3-code
- sort out columns without region or fin-region

Examples:

Table 2: Country <-> position in world map - ISO-Code

long	lat	Country	ISO3
-58.10884	75.20493	Greenland	GRL

long	lat	Country	ISO3
-70.48286	-52.00225	Argentina	ARG
104.00635	19.23091	Viet Nam	VNM
32.88457	-18.72852	Mozambique	MOZ
-79.14228	56.13643	Canada	CAN

```
## [1] "world map: 93248 rows x 12 columns"
```

3 Methods and Analysis

In this section the data sets from investigation section are combined, t3he countries are grouped by their:

- Region of the world
- Sub-Region
- income-level

and tried to evaluate patterns in the relation to the measures against the Covid-19. A textual sentiment analysis is done with the detail texts of the measures.

Different machine learning are then tested on these detail texts to detect the sentiments by these algorithms.

3.1 Fiscal-Data: Budget per country for Covid-19 measures in % of GDP and billion USD

To get an overview about what massive amounts of money we are talking here the following two tables:

Top countries with their budgets against Covid-19 as per IMF[8]
General Government = Central Government + State- + Local Government[9]

Table 3: Country: Top Budget sorted by % of GDP

	Country	GovLevel	BudgetGDP	BudgetUSDbn
58	United States	Central Government	25.5	5328.0
33	New Zealand	Central Government	19.3	40.4
57	United Kingdom	Central Government	16.2	440.1
3	Australia	General Government	16.1	219.3
47	Singapore	Central Government	16.0	54.5
26	Japan	General Government	15.9	800.8
8	Canada	Central Government	14.6	240.6
19	Germany	General Government	11.0	418.9
30	Mauritius	General Government	10.2	1.2
6	Brazil	General Government	8.8	126.4

Table 4: Country: Top Budget sorted by billion USD

	Country	GovLevel	BudgetGDP	BudgetUSDbn
58	United States	Central Government	25.5	5328.0
26	Japan	General Government	15.9	800.8
10	China	General Government	4.8	710.6
57	United Kingdom	Central Government	16.2	440.1
19	Germany	General Government	11.0	418.9
8	Canada	Central Government	14.6	240.6
3	Australia	General Government	16.1	219.3
17	France	General Government	7.6	198.6
25	Italy	General Government	8.5	159.8
6	Brazil	General Government	8.8	126.4

[8] "Fiscal Policies Database / Fiscal-Policies-Database-in-Response-to-COVID-19, International Monetary Fund (IMF)" (2021)
[9] "Definitions of Government in IMF-Supported Programs" (2013)

3.2 FCI-data: Measures: categories level 1 + 2 top scorer

Possible measures are defined by WBI/FCI[10] in a 5 x 19 matrix (=95 possible measures). The banking sector is the top scorer in Covid-19 policies and measures: (NBFI)[11]

Table 5: Top n measures level 1

MeasureL1	n
Banking Sector	2091
Liquidity/Funding	913
Financial Markets/Nbfi	488
Payment Systems	278
Insolvency	51

```
## [1] "2091 (= 54.7 %) of  3821 measures in banking sector!"
```

Currently only 12 measures with a **significant usage count > 30** are used, so there is potential to use more measures. in particular, measures for the providers of the financial systems seem to be significantly more than for the customer of the financial systems - i.e. measures against insolvency seem to be underrepresented.

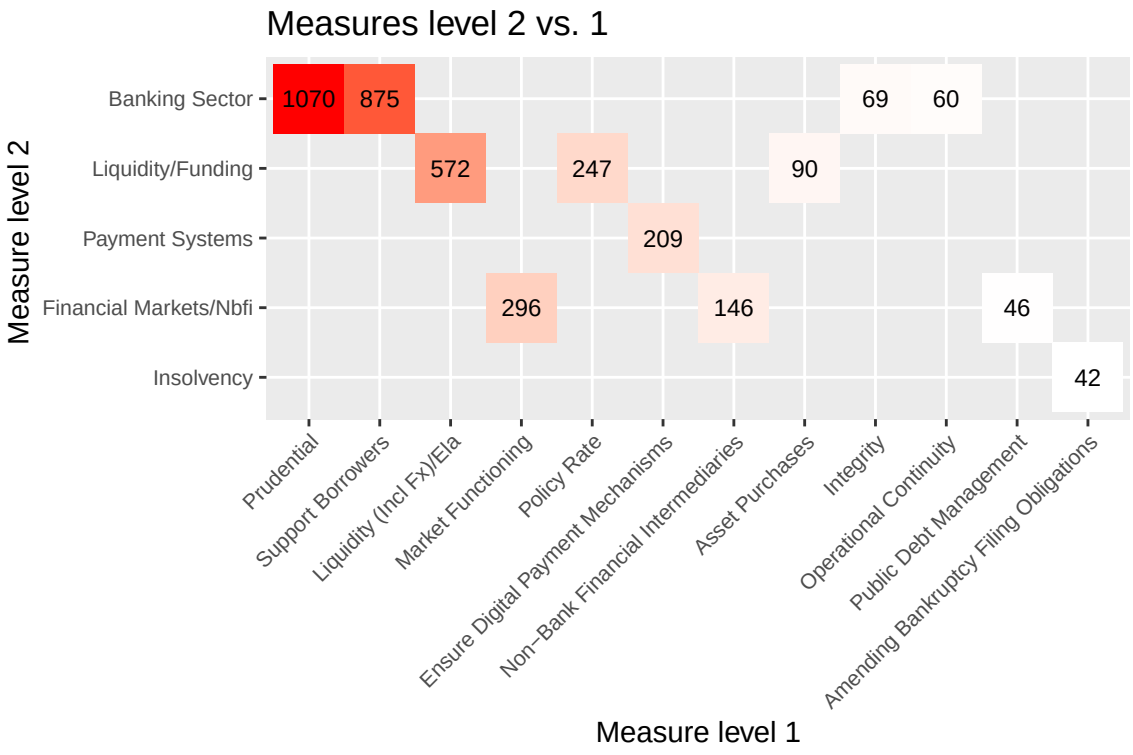


Figure 3: Number of measures level 2 vs. 1

[10] "COVID-19 Finance Sector Related Policy Responses Catalog" (2021);"COVID-19 Finance Sector Related Policy Responses Data Link / Covid-Fci-Data.xlsx" (2021)

[11] NBFI = Non-bank financial intermediaries, i.e. pension funds, insurance companies, mutual funds, venture capital

3.3 FCI-data: Textual sentiment analysis on the details

The approach is done as per chapter 26.3 course book[12] plus additional web resources.[13] [14]

The detail texts of countries are combined and analyzed in the following steps t_0 .. t_9:

- t_0 = filter needed columns out of FCI data
- t_1 = tokenize / group / count detail
- t_2 = remove stop words and numbers - not absolutely needed, but to get an overview
- t_3 = core of sentiment analysis = inner join to sentiments, by that all "noise words" are dropped out
- t_4 = cleanup NA, first plot of words
- t_5 = rejoin Country
- t_6 = filter out rare words
- t_7 = add Region
- t_8 = add SubRegion
- t_9 = add IncomeLevel <- this is the main result used in the later sections

Examples for measure details:

Detail
ESRB adopted varous recommendations: (iii) liquidity risks arising from margin calls.
Banks are now allowed to make maximum of four (4) restructuring for a credit exposures affected by the pandemic, irrespective of the number of times it has been restructured in the past.
The EBA communicated on 21 September the phaseâ€?out of its GL on moratoria. In the light of the second COVIDâ€?19 outbreak and the resulting government restrictions in many EU countries, the EBA has decided to reactivate the GL on moratoria by introducing a new deadline for the application of moratoria of 31 March 2021, replacing the previous date of 30 September 2020. These constraints aim to ensure that credit losses continue to be reflected in banks' balance sheets, including in relation to loans that are covered by the current 69 EU general payment moratoria that have been reported to the EBA
BFM announced that banks can deduct the amount of restructured credits since March 2020 from the mandatory reserves.
1) Pandemic Emergency Purchase Programme (PEPP) with an overall envelope of â,¬750 billion. Purchases will be conducted until the end of 2020 and will include all the asset categories eligible under the existing asset purchase programme (APP), 2) expansion of the range of eligible assets under the corporate sector purchase programme (CSPP) to non-financial commercial paper, making all commercial papers of sufficient credit quality eligible for purchase under CSPP, 3) easing the collateral standards by adjusting the main risk parameters of the collateral framework)

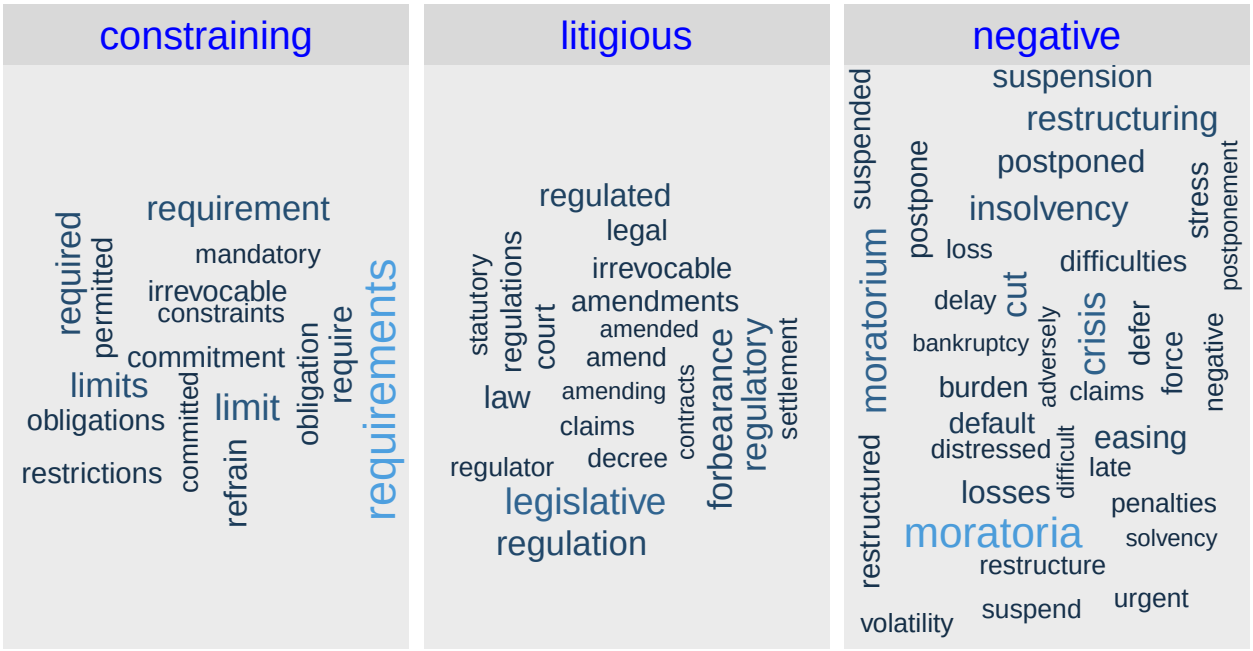
```
## Joining, by = "word"
## Joining, by = "word"
```

[12] https://rafalab.github.io/dsbook/
[13] Jan Kirenz, R Text Mining (2019)
[14] "Parsing Text for Emotion Terms Analysis Visualization Using R" (2019)

3.4 Sentiments: clouds and maximum sentiments

3.4.1 Sentiment clouds: Portfolio of sentiment types as wordclouds

5 main sentiment groups could be found in the detail texts.
The fill grade of the boxes reflect the frequency of the sentiment type, the size of the words their frequency.[15]



[15] E. Le Pennec (2020)

3.4.2 Sentiment type and maximum: per countries and country groups

The following heatmap is sorted in X- and Y-dimension to have the highest number of sentiment top left and the lowest bottom right. To simplify we choose the maximal number as representative for the country. We see the top sentiment is nearly 2 times the 2nd sentiment (Spain: 120 negative, 74 litigious, Italy: 105-44). In chapter 4.3 we'll verify this a step deeper with boxplots.

The following sentiment heatmap is sorted by X and Y dimension top left to low right. To simplify in further steps, we choose the maximum sentiment as a representative for the country. We see, the top sentiment is +/- 2 times the 2nd rank sentiment (Spain: 120 negative <-> 74 litigious, Italy: 105 <-> 44). In chapter 4.3 we will review in more detail with boxplots.

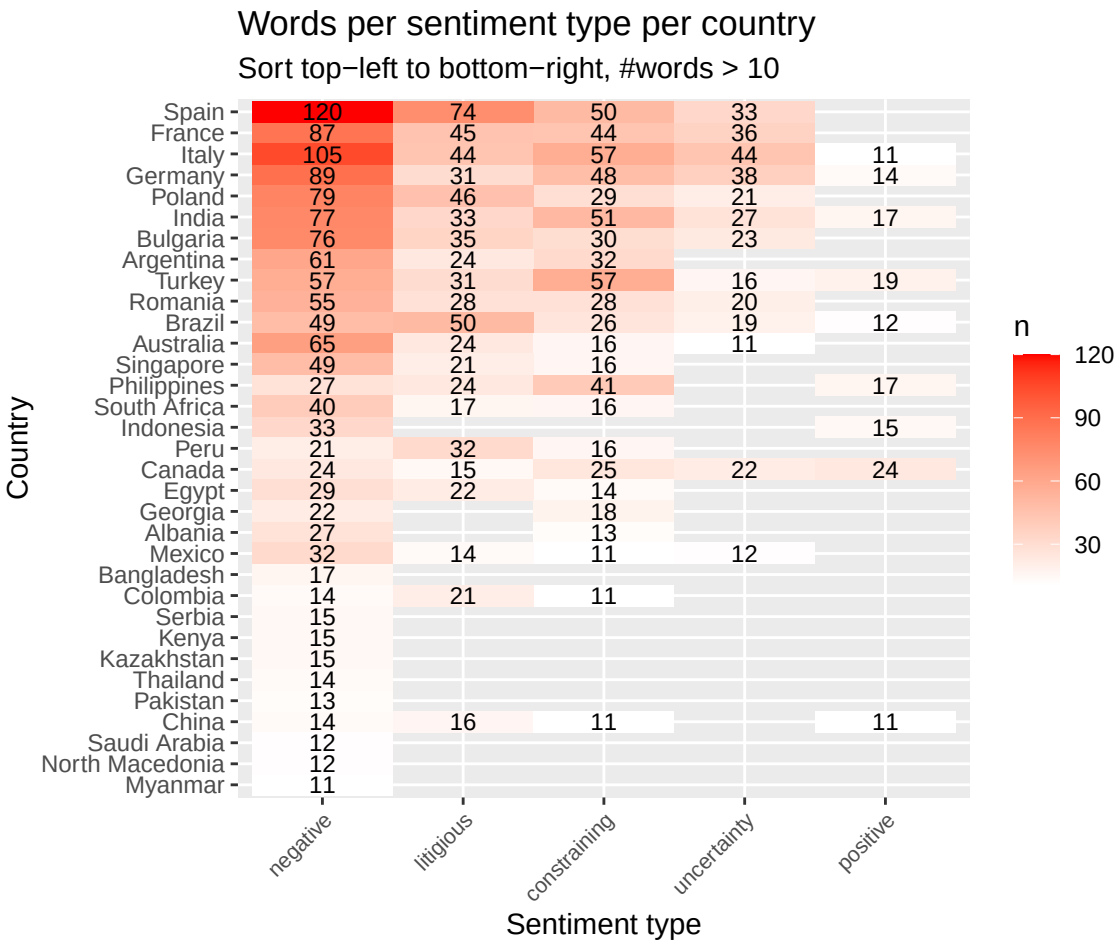


Figure 4: wordcloud portfolio

The table above shows an anomaly: although the litigious top words are in front of the constraining, in total the constraining are more than the litigious words (Rank 2 and 3 changes depending of perspective).

Words sentiment distribution over all texts:

Table 7: maximal sentiment counts over all words

negative	constraining	litigious	uncertainty	positive
2824	1494	1255	795	552

3.5 Machine Learning Approaches

The following algorithms were used and investigated and searched for optimal parametrization and speed:

- Naive-Bayes with no use of laplace parameter
- Naive-Bayes with search for optimal laplace parameter
- Fast Naive-Bayes multinomial and Bernoulli method
- Support Vector Machines
- K nearest neighbor manual search for k
- K nearest neigbour using the caret functionality to search the best k

3.5.1 Prepare special data structures for the machine learner: the corpus.

corpus = here interpreted as the measure detail texts (=documents) per country

Build up the corpora from the countries corpus and perform a cleanup: remove numbers, stopwords, punctuation, white space and perform word stemming to prepare optimal comparability:[16]

** 3728 'documents' of 152 countries = 16.6MB of text **

Three example corpus of the corpora:

```
## [1] "Extension of overdraft debt due between March-June until October and suspension of supervisory a
## [1] "   extens overdraft debt due marchjun octob suspens supervisori administr fee penalti overdraft
## [1] "Relaxation of provisioning requirements for loans that are covered by partial credit guarantees
## [1] "   relax provis requir loan cover partial credit guarante"
## [1] "Liquidity contingency plan in effect by DAB, including request for FSD notes to be brought\r\n i
## [1] "   liquid conting plan effect dab includ request fsd note brought countri arrang liquid provis l
```

3.5.2 Create the Document-Term-Matrix (DTM), remove sparse terms

For the following algorithms a needed input type

- rows = indicate the text
- columns = indicate the words
- cells = frequency of words in text

Remove those words from the DTM which are in "nearly all" documents - these do not help to cluster the words, called "remove sparse terms"[17]

[16] https://stackoverflow.com/questions/9637278/r-tm-package-invalid-input-in-utf8towcs
[17] https://stackoverflow.com/questions/28763389/how-does-the-removesparseterms-in-r-work

3.5.3 Visualize the cleaned / stemmed corpus

With the following wordcloud the resulted corpora with the words and their frequency (=size of the word) is shown. The more in the middle a word is located, the more frequent the word is occurring in the corpora.

Figure 5: Cleaned / stemmed corpus

3.5.4 Separate into train- (80%) and testset (20%)

Separate and check that the test and train data are representative = have similar distribution

Table 8: Frequency of test and train data are similar

train..	train.Freq	test..	test.Freq	delta_percent
constraining	0.0563241	constraining	0.0658762	-1.0
litigious	0.0830040	litigious	0.0843215	-0.1
negative	0.8474967	negative	0.8300395	1.7
positive	0.0029644	positive	0.0039526	-0.1
uncertainty	0.0102108	uncertainty	0.0158103	-0.6

3.5.5 Reduce the number of features

To complete our data preprocessing, we reduce the number of features in our test and training DT Matrices. To do this, we will use the findFreqTerms() function (again found in the tm package) which filters out all words below specific frequency - here freq=5 - minimal 5 times a word must occur to be not filtered out.

3.5.6 Naive-Bayes algorithm[18]

Now the data structures are prepared for the Naive-Bayes algorithm:

- convert the sparse Document-term-matrix from numeric to categorical "yes/no" matrices that the e1071-library-algorithm can process
- train model and check accuracy on test data (and measure time)
- use Naive-Bayes
  a) without laplace parameter
  b) with laplace parameter

3.5.6.1 Naive-Bayes algorithm without using laplace (= 0)

```
## Confusion Matrix and Statistics
##
##               Actual
## Predicted      constraining litigious negative positive uncertainty
##   constraining           18         8       61        2           0
##   litigious              11        32       89        1           3
##   negative               21        24      478        0           5
##   positive                0         0        0        0           0
##   uncertainty             0         0        2        0           4
##
## Overall Statistics
##
##                Accuracy : 0.7009
##                  95% CI : (0.667, 0.7333)
##     No Information Rate : 0.83
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.2516
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: constraining Class: litigious Class: negative
## Sensitivity                      0.36000          0.50000          0.7587
## Specificity                      0.89986          0.85036          0.6124
## Pos Pred Value                   0.20225          0.23529          0.9053
## Neg Pred Value                   0.95224          0.94864          0.3420
## Prevalence                       0.06588          0.08432          0.8300
## Detection Rate                   0.02372          0.04216          0.6298
## Detection Prevalence             0.11726          0.17918          0.6957
## Balanced Accuracy                0.62993          0.67518          0.6856
##                      Class: positive Class: uncertainty
## Sensitivity                 0.000000           0.333333
## Specificity                 1.000000           0.997323
## Pos Pred Value                   NaN           0.666667
```

[18] https://datascienceplus.com/sentiment-analysis-with-machine-learning-in-r/

```
## Neg Pred Value              0.996047          0.989376
## Prevalence                  0.003953          0.015810
## Detection Rate              0.000000          0.005270
## Detection Prevalence        0.000000          0.007905
## Balanced Accuracy           0.500000          0.665328
```

```
## [1] "   best parameter:0.00 - accuracy: 70.092%"
```

3.5.6.2 Naive-Bayes with laplace - some manual assumptions / trainings

Loop with 8 different laplace parameter values

```
## [1] "Naive-Bayes e1071 with laplace"
```

```
## [1] "        parameter:0.01 - accuracy: 70.619%"
## [1] "        parameter:0.02 - accuracy: 70.619%"
## [1] "        parameter:0.05 - accuracy: 70.751%"
## [1] "        parameter:0.10 - accuracy: 70.751%"
## [1] "        parameter:0.20 - accuracy: 70.619%"
## [1] "        parameter:0.50 - accuracy: 68.906%"
## [1] "        parameter:0.75 - accuracy: 71.014%"
## [1] "        parameter:1.00 - accuracy: 70.751%"
```

```
## [1] "   best parameter:0.75 - accuracy: 71.014%"
```

3.5.7 Two fast algorithms from Fast-Naive-Bayes-library used for the Document-Term-Matrix

Now we are evaluating the Document-Term-Matrices with the same laplace parameter as above but with 2 fast algorithms from Fast-Naive-Bayes-library[19]

Algorithms:

- Multinomial
- Bernoulli

Loop with 8 different laplace parameter values as above

```
## [1] "Naive-Bayes e1071 with laplace"
```

```
## [1] "        parameter:0.01 - accuracy: 75.494%"
## [1] "        parameter:0.02 - accuracy: 75.626%"
## [1] "        parameter:0.05 - accuracy: 75.362%"
## [1] "        parameter:0.10 - accuracy: 74.835%"
## [1] "        parameter:0.20 - accuracy: 75.099%"
## [1] "        parameter:0.50 - accuracy: 75.099%"
## [1] "        parameter:0.75 - accuracy: 75.099%"
## [1] "        parameter:1.00 - accuracy: 75.099%"
```

```
## [1] "   best parameter:0.02 - accuracy: 75.626%"
```

[19] https://cran.r-project.org/web/packages/fastNaiveBayes/vignettes/fastnaivebayes.html

```
## [1] "Fast Naive-Bayes multinomial"
```

```
## [1] "        parameter:0.01 - accuracy: 65.349%"
## [1] "        parameter:0.02 - accuracy: 65.349%"
## [1] "        parameter:0.05 - accuracy: 65.481%"
## [1] "        parameter:0.10 - accuracy: 64.954%"
## [1] "        parameter:0.20 - accuracy: 64.954%"
## [1] "        parameter:0.50 - accuracy: 64.163%"
## [1] "        parameter:0.75 - accuracy: 65.613%"
## [1] "        parameter:1.00 - accuracy: 66.140%"
```

```
## [1] "   best parameter:1.00 - accuracy: 66.140%"
```

3.5.8 Document-Term-Matrix as input for a Support Vector Machine (SVM)

These SVM shall be good for textual analysis:[20] [21]

```
## [1] "Support vector machines"
```

```
## Confusion Matrix and Statistics
##
##               Actual
## Predicted      constraining litigious negative positive uncertainty
##   constraining            0         0        0        0           0
##   litigious               0         0        0        0           0
##   negative               50        64      630        3          12
##   positive                0         0        0        0           0
##   uncertainty             0         0        0        0           0
##
## Overall Statistics
##
##                Accuracy : 0.83
##                  95% CI : (0.8014, 0.8561)
##     No Information Rate : 0.83
##     P-Value [Acc > NIR] : 0.5235
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: constraining Class: litigious Class: negative
## Sensitivity                      0.00000          0.00000            1.00
## Specificity                      1.00000          1.00000            0.00
## Pos Pred Value                       NaN              NaN            0.83
## Neg Pred Value                   0.93412          0.91568             NaN
## Prevalence                       0.06588          0.08432            0.83
## Detection Rate                   0.00000          0.00000            0.83
## Detection Prevalence             0.00000          0.00000            1.00
```

[20] https://rpubs.com/masmit11/533879
[21] https://stackoverflow.com/questions/40051542/text-classification-using-e1071-svm

```
## Balanced Accuracy                0.50000          0.50000            0.50
##                       Class: positive Class: uncertainty
## Sensitivity                  0.000000            0.00000
## Specificity                  1.000000            1.00000
## Pos Pred Value                    NaN                NaN
## Neg Pred Value               0.996047            0.98419
## Prevalence                   0.003953            0.01581
## Detection Rate               0.000000            0.00000
## Detection Prevalence         0.000000            0.00000
## Balanced Accuracy            0.500000            0.50000
```

```
## [1] "   best parameter:         accuracy: 83.004%"
```

3.5.9 Document-Term-Matrix as input for K-Nearest-Neigbur algorithm (KNN):

"When using KNN for classification, it is best to assess odd numbers for k to avoid ties..."[22]

Default value is: one neighbor - we test with 1/3/5/../15 neighbors - as above also 8 loop counts.

```
## [1] "K nearest neighbor loop approach"
```

```
## [1] "neighbours:01 - accuracy: 79.315%"
## [1] "neighbours:03 - accuracy: 82.872%"
## [1] "neighbours:05 - accuracy: 82.609%"
## [1] "neighbours:07 - accuracy: 82.872%"
## [1] "neighbours:09 - accuracy: 82.740%"
## [1] "neighbours:11 - accuracy: 82.872%"
## [1] "neighbours:13 - accuracy: 82.477%"
## [1] "neighbours:15 - accuracy: 82.740%"
```

```
## [1] "   best parameter:3.00 - accuracy: 82.872%"
```

3.5.10 Optimize k (neighbors) with the caret library

Finally use the caret library to evaluate automatically the optimal parameter k (number of neighbors). Here the Document-Term-Matrix has to be slightly adapted into binary form - instead of "yes"/"no" for KNN here it needs 0/1.

```
## [1] "K nearest neighbor caret optimizer"
```

```
## k-Nearest Neighbors
##
## 3036 samples
##  443 predictor
##    5 classes: 'constraining', 'litigious', 'negative', 'positive', 'uncertainty'
##
## Pre-processing: centered (443), scaled (443)
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 2733, 2732, 2733, 2733, 2732, 2733, ...
## Resampling results across tuning parameters:
```

[22] https://bradleyboehmke.github.io/HOML/knn.html

```
##
##   k   Accuracy   Kappa
##    5  0.8478419  0.19728342
##    7  0.8501575  0.12530007
##    9  0.8508057  0.07899331
##   11  0.8511237  0.05851006
##   13  0.8514538  0.05499845
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 13.
```

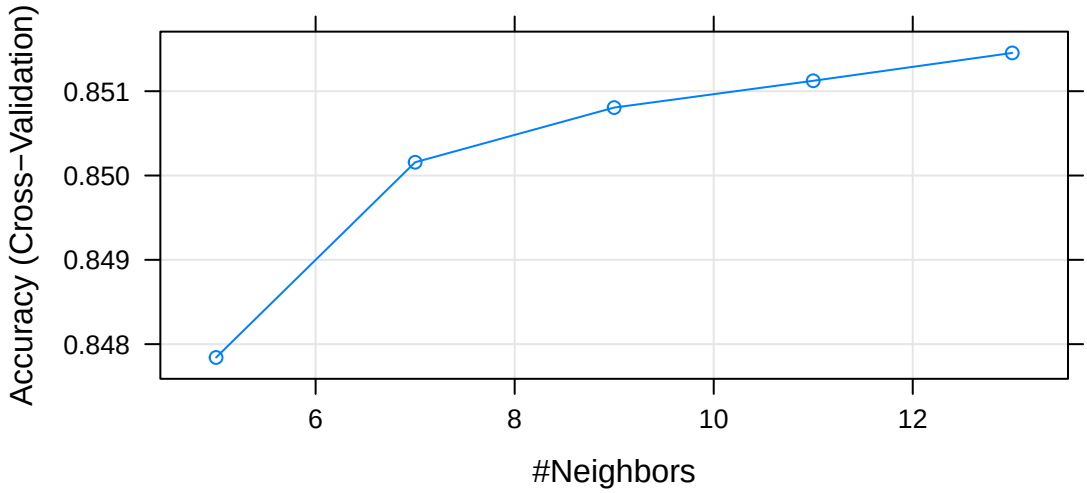


Figure 6: knnFit-Results: optimal k value

4 Results

4.1 Maximum sentiment per country

As shown before the policies and measures related to Covid-19 with **negative sentiment** is prevailing, followed by constraining and litigious sentiments.
Countries using mainly measures with uncertain or positive sentiments could nearly not be seen from this high perspective.
Policies and measures are not defined for all countries in the FCI-data (only for 156 of them).
These countries are greyed out in the following map ("NA").

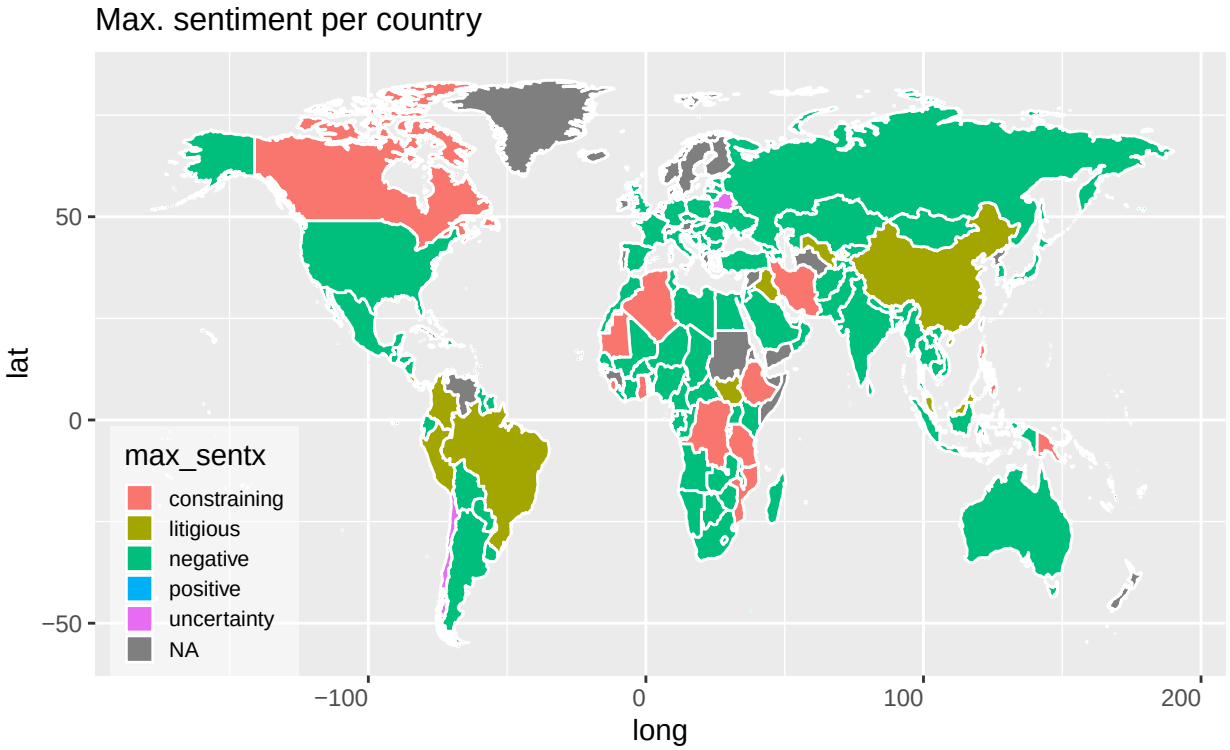


Figure 7: Max. sentiments per country

Here as table the distribution on sentiment level per country:

Table 9: Maximal sentiment counts of countries

max_sentx	count
constraining	16
litigious	10
negative	120
positive	1
uncertainty	2

Table 10: Very few Countries with mainly positive or uncertain measures

Country	max_sentx
Belarus	uncertainty
Chile	uncertainty
Jamaica	positive

4.2 Sentiments, country groups and boxplots

The following diagrams show the word count of sentiment type per region, income level, and financial region. This shall demonstrate that the maximum sentiment of the countries can be considered as representative for the countries measure, the height and width of the more left boxes is more than that on the right side. Example: top chart (per Region), left sentiment (negative), blue box (in Europe): ~25-80 negative words, median at ~62. The other blue boxes are much smaller, median around 25.

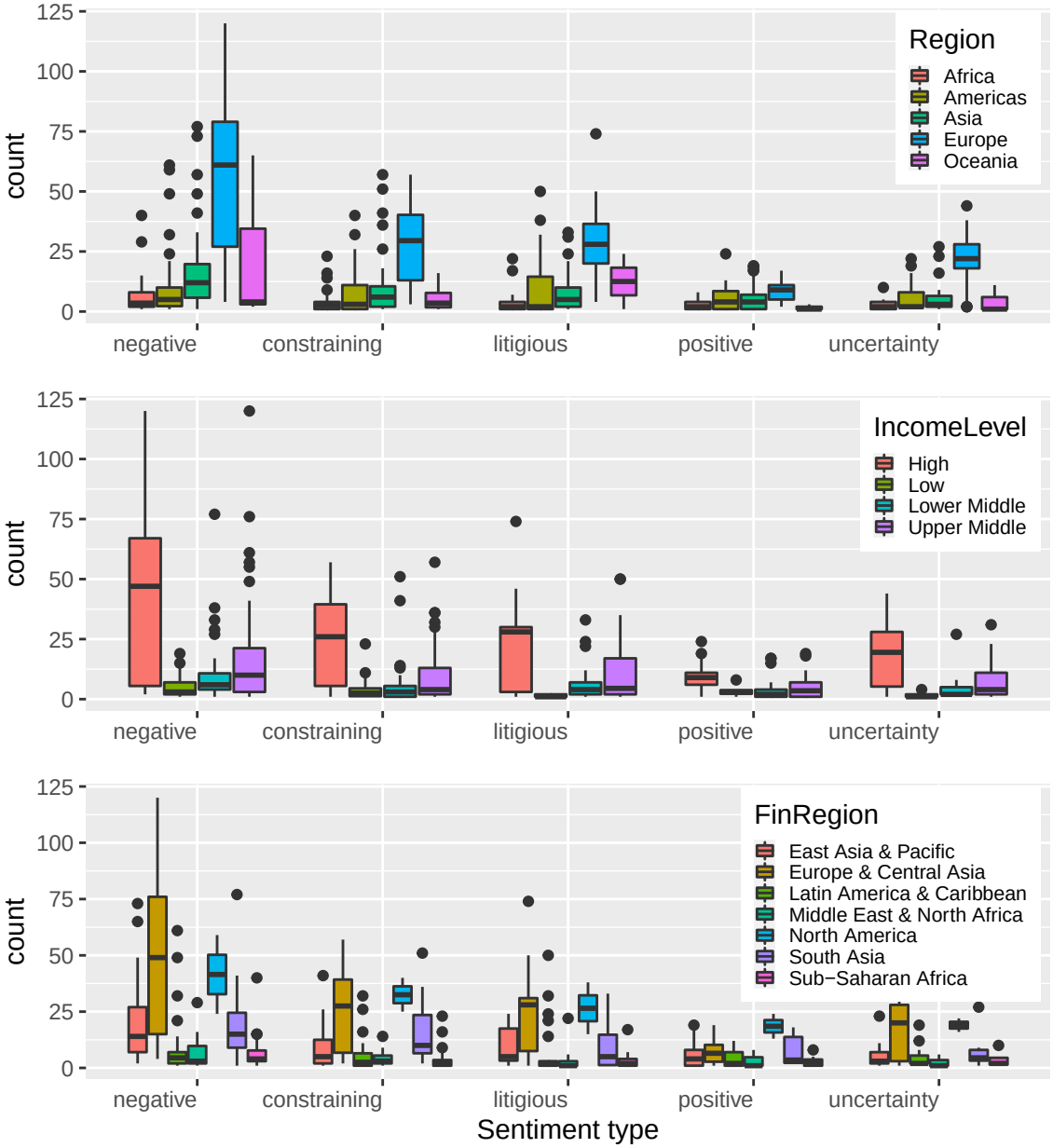


Figure 8: Sentiment types per country group

4.3 Machine Learning Ratings

The following results were achieved during using different machine learning methods.

One can see the optimal parameter of the algorithms to get the shown accuracy within the shown time (in seconds).

The best overall accuracy between these investigated algorithms can be achieved by the KNN method optimized by the caret library.

A 100 times faster approach with only 2% less accuracy is the support vector machine method.

Type of Machine learning	best parameter	Accuracy	timing(sec)
Naive-Bayes e1071 no laplace	0	70.09	10.43
Naive-Bayes e1071 with laplace	0.75	71.01	82.78
Fast Naive-Bayes multinomial	0.02	75.63	0.44
Fast Naive-Bayes Bernoulli	1	66.14	0.43
Support vector machines	0	83	3.02
K nearest neighbor loop approach	3	82.87	167.12
K nearest neighbor caret optimizer	13	85.15	286.03

5 Conclusion

The research questions of chapter 2 could be answered as follows:
Policy measures in the financial world related to the Covid-19 pandemic could be shown by budget[23] and measure level. Only a small subset of possible measures are activated.[24]

The policies and measures sentiments are mainly negative, followed by constraining or litigious measures. With the word clouds we showed what these sentiment words mean in the financial world. Nearly no countries have mainly positive (1 country = Jamaika) or unsecure (2 countries) measures.[25]

Common measures, regardless of region and income level are strongly focused on the banking sector - two of their measures cover 55% of all measures. Measures for the providers of the financial systems seem to be significantly more than for the customer of the financial systems - i.e. measures against insolvency seem to be underrepresented[26]

Machine learning algorithms can be used to predict the sentiments of measures for the countries. The KNN algorithm delivers the most exact result, fast and slightly less exact is the SVM algorithm.[27]

As next step one could investigate more in the grouping of the countries, to find a better grouping criterion for a pandemic beside fiscal or regional aspects.

Thank you for reading, I hope you have enjoyed it as much as I have enjoyed the exploring.

[23] Chapter 3.1
[24] Chapter 3.2
[25] Chapter 3.4
[26] Chapter 3.2
[27] Chapter 3.5

6 Appendix

6.1 Abbreviations

Abbreviation	Explanation
bn	billion
DTM	Document-Term-Matrix
FCI	Finance, Competitiveness & Innovation defined by Worldbank
FRED	Federal Reserve Economic Data (FRED)
KNN	K-Nearest-Neighbour machine lerarning algorithm
NBFI	Non-bank financial intermediaries, i.e. pension funds, insurance companies, mutual funds, venture capital
SRAF	Software Repository Accounting and Finance, University of Notre Dame, IN, USA
SVM	Support vector machines machine lerarning algorithm
WBI	Worldbank Institute

6.2 References

"Countries with Regional Codes / All.csv." 2021. https://raw.githubusercontent.com/lukes/ISO-3166-Countries-with-Regional-Codes/master/all/all.csv.

"Country and Lenting Group of WBI - Knowledge Base." 2021. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups.

"COVID-19 (Coronavirus) Response." 2021. https://www.worldbank.org/en/topic/health/coronavirus.

"COVID-19 Finance Sector Related Policy Responses Catalog." 2021. https://datacatalog.worldbank.org/dataset/covid-19-finance-sector-related-policy-responses/.

"COVID-19 Finance Sector Related Policy Responses Data Link / Covid-Fci-Data.xlsx." 2021. https://development-data-hub-s3-public.s3.amazonaws.com/ddhfiles/936261/covid-fci-data.xlsx.

"Datanovia, HOW TO CREATE A MAP USING Ggplot2." 2021. https://www.datanovia.com/en/blog/how-to-create-a-map-using-ggplot2/.

"Definitions of Government in IMF-Supported Programs." 2013. https://www.imf.org/external/pubs/ft/tnm/2013/tnm1301.pdf.

"Documentation Loughran McDonald MasterDictionary (SRAF)." 2018. https://www3.nd.edu/~mcdonald/Word_Lists_files/Documentation/Documentation_LoughranMcDonald_MasterDictionary.pdf.

E. Le Pennec. 2020. "Ggwordcloud, a Word Cloud Geom for Ggplot2." https://lepennec.github.io/ggwordcloud/articles/ggwordcloud.html.

"Fiscal Policies Database / Fiscal-Policies-Database-in-Response-to-COVID-19, International Monetary Fund (IMF)." 2021. https://www.imf.org/en/Topics/imf-and-covid19/Fiscal-Policies-Database-in-Response-to-COVID-19.

"International Monetary Fund (IMF)." 2021. https://www.imf.org.

"International Monetary Fund (IMF) COVID-FM-Database." 2021. https://www.imf.org/en/Topics/imf-and-covid19/~/media/Files/Topics/COVID/FM-Database/SM21/revised-april-2021-fiscal-measures-response-database-publication-april-2021-v3%22.

Jan Kirenz, R Text Mining. 2019. "R Text Mining." https://www.kirenz.com/post/2019-09-16-r-text-mining/.

"Our World In Data (OWID)." 2021. https://www.owid.org.

"Parsing Text for Emotion Terms Analysis Visualization Using R." 2019. https://datascienceplus.com/parsing-text-for-emotion-terms-analysis-visualization-using-r-updated-analysis.

R.A. Irizarry. 2021. "Introduction to Data Science." https://rafalab.github.io/dsbook/.

"Software Repository Accounting and Finance / LM Sentiment Word Lists, University of Notre Dame (IN,USA)." 2018. https://sraf.nd.edu/textual-analysis/resources/#LM%20Sentiment%20Word%20Lists.

"Software Repository Accounting and Finance (SRAF), University of Notre Dame (IN,USA)." 2021. https://sraf.nd.edu/textual-analysis/resources/.

"Software Repository Accounting and Finance / Textual Analysis / Resources, University of Notre Dame (IN,USA)." 2018. https://sraf.nd.edu/textual-analysis/resources/.

T.Loughran, B.McDonald. 2010. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." https://poseidon01.ssrn.com/delivery.php?ID=496096085110072018116093082122093101099083089080001063109
EXT=pdf&INDEX=TRUE.

"The World Bank Group's Response to the COVID-19 (Coronavirus) Pandemic." 2021. https://www.worldbank.org/en/who-we-are/news/coronavirus-covid19.

"Tidyquant - Bringing Financial and Business Analysis to the Tidyverse, Daily Stock Prices (and Many Other) from Federal Reserve Economic Data (FRED)." 2021. https://business-science.github.io/tidyquant/.

"World Bank List of Economies (June 2020), Country and Lenting / CLASS.xls." 2021. http://databank.worldbank.org/data/download/site-content/CLASS.xls.

"Worldbank Institute (WBI)." 2021. https://www.worldbank.org.

Yihui Xie, Christophe Dervieux, Emily Riederer. 2021. "R Markdown Cookbook." https://bookdown.org/yihui/rmarkdown-cookbook/.