Case Study of Big Data Methods for Innovation Detection in the Netherlands
Drop in Classification Model Accuracy Investigation

Locusteanu, Eva Maria

Award date:
2019

Case Study of Big Data Methods for Innovation Detection in the Netherlands

Drop in Classification Model Accuracy Investigation

Locusteanu Eva Maria
e.m.locusteanu@student.tue.nl
Student Number 1036088

A thesis presented for the degree of
Bachelor of Science

Department of Mathematics and Computer Science
Eindhoven University of Technology
Netherlands
July 24, 2019

Abstract

Big data sources, in combination with traditional collection methods such as surveys can be an advantage in monitoring many indices in the country. Big data has been shown to improve the process of compiling national statistics, by making it faster, cheaper and by allowing for interesting unexpected by-products. At Statistics Netherlands this approach has been tried out, by deploying a model that detects whether a company is innovative or not based on its website. Now, since the original model performed well with an accuracy of 92 % it has been surprising to see that the performance deteriorated with every new try, reaching its lowest point on the fifth collected dataset with an accuracy of 60 %. I investigate the decrease in accuracy of the innovation detection model using NLP approaches and statistical inference. I furthermore design a project pipeline theory that matches the object of study and in the end evaluate, based on my analysis, the potential risk zones that are most likely related to the differences in accuracy.

Keywords: big data; survey statistics; innovation; natural language processing

1 Introduction

National Institutes of Statistics (NSIs) normally employ two types of data in their procedures: survey data, or primary data and administrative data, or secondary data. Key reasons such as increased speed, efficiency, spatial granularity have motivated NSIs in recent years to incorporate big data into their data sources as well [7] [6]. (For a tabular comparison of these three types of data see Appendix A.) There are in principle three major ways in which big data can benefit macroeconomic and financial statistics: (1) by answering questions and producing new indicators; (2) by bridging time lags in the availability of official statistics and supporting the timelier forecasting of existing indicators and (3) as an innovative data source in the production of official statistics. [13].

A classification of big data sources divides these into three categories: big data can stem from (1) social networks (human-produced), (2)traditional business systems (process-based) and (3) internet of things (machine-generated) - see Appendix B for ways in which these three categories relate to primary or secondary data, and elaborate discussions in [13].
Promising examples of big data sources from all of these fields that have been used for macroeconomic purposes are various. They include for instance the use of mobile phone network data to estimate populations of small areas . The Office for National Statistics in the UK and the Central Statistics Office in Ireland are experimenting with the analysis of internet search queries within migration statistics [7], and the use of electricity smart meter data to determine household composition [4]. Research from Telefonica [8] suggests that mobile phone records can be used for forecasting socio-economic trends as well as predicting socio-economic levels of a population. Even more, an Eurostat feasibility study provides valuable insights into the use of mobile positioning data for population statistics, and tourism information [18]. Emilio Zagheni estimated global migration trends by analysing 43 million anonymous Yahoo! account holders IP addresses as stated in [7].Social media and query data have further formed basis of optimistic examples of estimating macro-economic indicators : Google Web searches and Facebook posts have been successfully used to predict stock market liquidity or to construct sentiments metrics that predict stock market activity [2] [14].

1.1 Present study: Innovation Classification Model

Building on the trend NSIs have seen of accommodating big data sources in building national statistics [13] the Centre for Big Data Statistics of Statistics Netherlands(CBS) conceived a novel way to detect innovation in Dutch companies has been proposed. Normally, the way CBS collects data about innovative companies is via a survey, that gets sent out every other 2 years to each company in the country with more than 10 employees. This survey measures only product technological innovation. The surveying method takes a long time and is expensive. Therefore a new way to detect innovation that does not depend entirely on the survey. Instead of questionnaire outputs, the data proposed to be analyzed is the text on a company's website.

The **purpose of the model** satisfies two of the three main benefits of big data in improving large scale statistics, as it (1) satisfies potential of big data to act as an innovative data source in the production of official statistics on one hand by expanding the amount of data available of innovation profiling in Dutch companies, but also on the other hand by answering new questions (2). The latter refers to the question of small innovative companies, defined as enterprises with less than 10 employees. For this subset, CBS collects no primary data and therefore no innovation indicators are available via traditional methods. This particular small companies group is especially interesting to monitor as it measures/relates to the overall amount of start-ups in the country.

The model employed by CBS has as input data text scrapped from websites. All of the websites have labels (1 = innovative, 0 = not innovative), based on the latest official survey results. Therefore a supervised learning prediction model was applied that has been trained to classify whether, based on the text on it's website, a company is innovative or not. Note that the notion of innovation for the purpose of this paper is referring to

the extent to which companies promote themselves as innovative, in a realistic sense. However, we assume that this proxy is good enough for the real measurement of innovation.

Furthermore a logistic regression model applied originally to dataset 1 gave the promising result of 92 % accuracy. After seeing promising results of the model on the first dataset, a replication of the model was tried on another slightly larger dataset (dataset 3, see Table 1) and with an improved scrapping method. However, the performance had drastically decreased. Similarly in order to understand the reason behind the decrease, dataset 4 was collected, this time with an even better scraping method. Since the accuracy continued to drop, suspicions arose with respect to the scraping method: that's why the original method was implemented again, as seen Table1. My project has started with these four datasets and my task was to understand the reason behind the incremental drops in accuracy. Note that I do not (and did not) have access to the second dataset.

Dataset	Size	Scrapping	Date of collection	Acc.(%)
Dataset 1.	4880	Collected by Suzanne van der Doef scraped with Beautiful soup;	1/08/2017	92
Dataset 2.	500.000	ABR data (websites from business register)	?	89
Dataset 3.	5391	Collected by Suzanne with Phantomjs	15/02/2018	75
Dataset 4.	5757	Collected by Piet Daas with Selenium server	21/01/2019	60
Dataset 5.	4674	Collected by Piet Daas and scraped with the original Beautiful soup approach	2/05/2019	60

Table 1: Datasets description.

The decrease in accuracy seen in Table 1 constitutes the object of my paper. Based on similarity metrics and statistical inference, I will answer this **research question** by narrowing down the potential causes behind this phenomenon.

Section 3 describes some of the theoretical background, both in terms of methodology and big data in general, but also theoretical background about NLP methods used in the methodology. Sections 4 is the analysis that refers to all four datasets (including methodology and results) and section 5 presents the analysis (methodology and results) pertaining only to differences between datasets 1 and 5 - choice justified in the following pages. Section 6 concludes with discussion points and interpretation of the results.

2 Pipeline

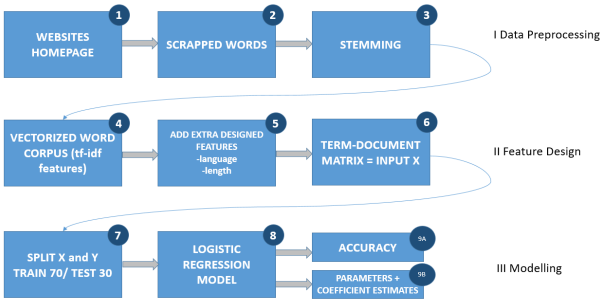

Figure 1: Model pipeline.

Having described the purpose of this model in the introduction, I now formulate the global **pipeline of the project**. In the first phase, data preprocessing lay the original websites homepages (1), transformed by scrapping into a text corpus (2) and then stemmed (3). Stemming is the procedure with which words such as "innovation" and "innovative" are mapped to the stem of the word, i.e. "innovat", for the purpose of NLP models generalizability. Next, in the feature design phase, the stemmed words are vectorized - in the present case using TF-IDF - (4), some additional relevant features such as language and document length are added (5) and then all of these are concatenated into a term-document matrix (6). In this matrix, each row represents a document and each column is a feature that refers to either individual terms in the document (as vectorized by the TF-IDF mechanism) or the additional features from the previous step. The columns and its contents per document are intrinsically related to the vectorization mechanism. Lastly this term-document matrix serves as input to any predictive model, and in our case we first split it into training and testing (7), then apply the logistic regression model (8) and finally extract the output: accuracy (9A) and parameters with their coefficients (9B).
This pipeline (Figure 1) will become relevant after presenting the analysis results, as I will point out where exactly the accuracy drop begins/can be explained.

3 Theoretical Background

3.1 Big Data & Pitfalls

As stated in the introduction, promising research initiatives from NSOs from Ireland and UK have shown that what used to be core population statistics topics such as migration and usual residents, which are in general

difficult to measure, can now be estimated using big data sources [7]. Even more examples can be given here. In Estonia, because Statistics Estonia has stopped using border surveys, a new method that used anonymized roaming mobile positioning data has been successfully developed to estimate international traveling information. Again migration patterns, normally difficult to estimate using primary data have been improved by speed, costs, accuracy and granularity by using big data sources.

Nonetheless big data pitfalls are just as present as the benefits. A well-known example of a big data pitfall is the case of Google Flu. Based on search queries, initially Google engineers were able to predict influenza disease prevalence in United States, at a granular and accurate level, even ahead of the official Centre for Disease Control (CDC) reports. [15].

Practically the vision of Google Flu was an admirable one - to paint a more accurate picture of the prevalence of a contagious disease that might eventually allow for life-saving interventions, suggesting again that big data has great potential and value can be extracted from it [15]. The way they formalized this vision - producing the flu estimates - was by matching a high number of queries (around 50 million search terms) to 1152 data points [9]. In doing so, many terms that matched the prevalence of the flu at a certain time but where structurally unrelated to it became very likely [15]. This is even more sustained by the choice of the Google Flu Trends (GFT) developers to randomly remove such correlations, technically unrelated to the flu - for example, queries related to the basketball season [9]. As a consequence, before 2009, their algorithm was both flu detector and winter detector [15]. Overestimating the flu effect however occured even after retouching the algorithm in 2009. It appears that the GTF flue estimates in 2011-2012 season had been more than twice than the correct CDC estimates [15]. In fact starting August 2011, GFT produced estimates consistently higher than those of CDC (overestimating the flu prevalence 100 out of 108 weeks) [15].

The reasons for the severe overestimation of flu prevalence by GFT are twofold [15]. On one hand, GFT developers were subject to what is called "big data hubris" - the (often implicit) assumption that big data sources can entirely replace, rather than supplement traditional statistical analysis methods [15], Indeed a model based on solely CDC reports lagged 2 weeks (think: small data) was able to produce comparable accuracy to the highly complex multi-milllion-query-based model from GFT [10]. Moreover, even after 2009 after the GFT algorithm had been corrected, only 3 weeks lagged CDC data performed better at predicting flu prevalence than the big data model [10], as follows:

$$\boldsymbol{flu}_t = \beta_0 + \beta_1 * \boldsymbol{flu}_{t-2} + \beta_2 * \boldsymbol{flu}_{t-3} + \epsilon$$

According to [10] a correlation of 0.86 was found between current flu levels and recently reported flue levels, using the autoregressive model above that only has lag t-2 and t-3. The first lag is missing because CDC provides their numbers with a two week delay instead of one. However, CDC has been working on a way to report flu levels faster in just one week and remodeling this using AR(3) [10] found the correlation to be 0.95 between current and recent flu levels, which compared to the GFT correlation of 0.94 is shockingly close. This is to say that when monitoring the flu, search data is comparable in utility to alternative information sources, but not necessarily superior. For an extensive discussion and more examples of big versus small data predictive capabilities in other domains as well such as movies, video games and music, please refer to [10]. And even more interesting than the comparative performance of big data sources to alternatives, is that a model that includes both is likely to outperform on all counts. [10](Figure 5) [15].

On the other hand GFT was subject to algorithmic dynamics - which is understood as the volatile relationship between the real-life construct to be measured (innovation in Dutch companies, or influenza-infected rates in the US) to the big data source chosen (web search, website text, tweets, etc.). To give an example, the search algorithm has been edited and reformulated by Google at least 86 times over the course of GFT [15]. Intuitively each change creates a discordance of the mapping between the real influenza prevalence and the way it is reflected eventually in the data that served as input to the model. I stress that other data sources are not intrinsically as dynamic. For instance in the case of mobile roaming data in Estonia to measure tourism economy and migration trends, we can imagine that a nonresident in the network of mobile operators is always going to be associated with an inbound traveler (and vice-versa for outbound travelers). Only perhaps in cases when a person is associated with multiple SIM cards, or phones are stolen etc. is this function likely to fail, but these cases occur relatively seldom, thus may not change the performance that much. So mobile data in this case measures the same thing across time - mobile data is said to have a stable predictive/informative power with respect to migration.

But beyond big data hubris and algorithmic dynamics, what are the challenges of analyzing big data, in general? Given the key features of big data: large sample size and heterogeneity [6] it is inferred that specific issues arise when analyzing these types of data, both in the domain of statistical inference as in the domain of storage and computing capacities. I only state the issues for the former: noise accumulation, spurious correlation, and incidental endogeneity are all very common to occur in big data analysis ; a discussion and proposed very technical solutions can be seen in [6].

The subdivision of challenges associated with big data (statistical inference and computation capacity) brings me to the next part of this section, where I describe how this division can be seen and understood from a different perspective.

3.2 Two cultures in data science: Big data before it was big

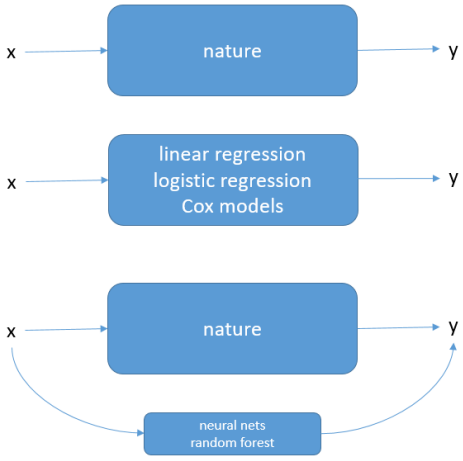

Figure 2: (top) General representation of data in statistics; (middle) Model-based approach; (bottom) Algorithmic approach. Adapted from [3]

Statistics is a field that intrinsically starts with data. In general, data is said to be generated by a black box, where a vector of input variables **x** go in on one side, and a vector of response variables **y** come out on the other side (see Figure 1 (top).)
There are two ways in which those concerned with the analysis of data perform their object of study, i.e. treat the black box character of nature [3]. On one hand, there is the model-based approach, in which one assumes that the data is generated by a stochastic data model [3]. The other approach can be related to computer science today, in which algorithmic models are used and the data mechanisms are treated as unknown [3]. In the **data modeling culture**, the start is usually an assumption that a stochastic model exists and is representative of the black box. Then a common such data model is that data points are generated by independent samples from $response\quad variables = f(explanatory\quad variables, random\quad noise, parameters)$. Parameter values are estimated and consequently the estimates are used for prediction purposes or other (see Figure 1(middle)). Then to evaluate the model, an "is this reasonable enough" - type question is answered, with options to answer being either yes or no, using goodness of fit methods and residual analyses. This type of modeling performed according to Breiman by roughly 98 % of statisticians.
With the **algorithmic modeling approach** on the other hand, no assumptions are made on the contents on the black box - the data process is in this case unknown. Within this culture, the task is to find a function $f(x)$ that maps predictors **x** to responses **y**. Validation is then done by predictive accuracy. A smart guess estimates the population of this type of approach among statisticians at 2 %.
Leo Breiman argues in his seminal paper "Two Cultures" that the focus in the statistical community on the data model has led to irrelevant theory and questionable scientific conclusions, that it has kept statisticians from using more suitable algorithmic models and even prevented statisticians from working on exciting new problems [3] .

To note is that examples of the algorithmic approach are based on genetic and astronomic data in [3]. Even though currently big data is understood by the three V's definition - high volume, high velocity and high variety - Breiman predicted the specific potential and challenges associated with this type of data well ahead of its burst in popularity caused by Internet 2.0. The fact is, the term big data now has a much wider and more relevant interpretation as it includes more domains such as social media, websites, internet transaction, sensor data, etc. What I will argue in this introduction is that, despite the broadened scope and relevance of big data, the distinction in data analysis attitudes presented by Breiman almost two decades ago still exists today. Even more so, I will show why precisely due to the increased scope of big data today, this distinction should bear a higher weight than it did at the beginning of the century: now practically everybody is involved with big data - individuals, business and institutions alike - not just a handful of astronomers and geneticists.

The ubiquitous use and production of data by business has attracted tremendous amount of attention to big data. This is supported by coining the term "data science", by creating job positions named "data scientist' and even by compiling university programs that educate students in how to become data scientists. However, it is clear that this increase of focus on big data stems from industry realizing there is value to be extracted from data analysis. All of a sudden, statistics is an interesting field, and analyzing data is no longer designed exclusively for statisticians but for everyone who is interested in improving their business

And yet, however hopeful it is that big data receives it's deserved attention both in industry and academia, this is not the whole truth. Graduate data science degrees

nowadays receive in equal amounts students with degrees in either Computer Science (representative of algorithmic school of thought) or Econometrics (data model culture) Similarly, positions entitled "data scientist" are filled by computer scientists and econometricians, the latter being in the majority. Breiman's distinction still exists, only better disguised.

The reason why this is a problem (that two distinct cultures are still present under the name of data science) is that there are very few people that understand both. The people who understand every single step from A to Z in a big data project, from scrapping and databases and cloud computing to statistical inferences and distributions and p-values are generally quite young and definitely in the minority. And because there are very few people who understand everything from A to Z, there is (except for very recent attempts from [7]) a lack of theory on a clear official methodology for big data application within traditional statistics.

I hope this research will add more knowledge about this methodology in the field, especially practical since the present research is based on a real life example (As opposed to theory alone as in [7].)

The last part in the theoretical background section will provide more insight into exactly how I plan to do this, i.e. using similarity measures to compare and asses differences between datasets, in order to hopefully point out if and where things go wrong in the pipeline.

3.3 Similarity Metrics

There are several ways to compare texts. Metrics imply the existence of a function that takes as parameters two vectors and outputs a number which typically falls into a certain range.CITE
Amongst ways to distinguish metrics of text similarity three large categories were proposed, as follows: string-based measures, corpus-based measures and Knowledge-based measures [11] [16].
String similarity measures operate on string sequences and character composition. A string metric is a metric that measures similarity or dissimilarity (distance) between two text strings for approximate string matching or comparison [11].
String-based measures can be further classified into two sub-categories: character-based and term-based similarities. Corpus-based measures of word semantic similarity try to identify the degree of similarity between words using information exclusively derived from large corpora. Lastly, knowledge-based measures use networks of meanings between words and contextual information to detect similarities [11].
In this paper, four similarity metrics will be used to answer the research question. All of them fall into the category of term-based metrics.
Cosine Similarity. Mostly Cosine distance metric is used to find similarities between different documents. This metric measures the degree of angle between two documents/vectors. Especially cosine similarity is useful when the magnitude of vectors is irrelevant, and instead the orientation is important. As opposed to euclidean distance, cosine similarity removes the effect of document length when comparing two documents.
Earth Mover Distance denotes the effort that needs to be made in order to move one ordering into another, also known as Wasserstein distance.
Jensen-Shannon Distance measures similarity between two distributions; being a metric, it has a range between 0 and 1, where 0 means that the two distributions are identical.
Hellinger Distance functions with a similar approach as the Jensen-Shannon distance, only it is not bounded.

4 Analysis I

4.1 Methodology

In order to answer the research question,i.e. is the difference in accuracy related to the difference in text, I focus take several steps. First, I explored the output of the model on each of the 4 datasets in order to gain a global understanding of what might be the cause behind the accuracy difference. The learning curves of the models were plotted for each case, as seen in Figure 3. Furthermore the output of the logistic regression model was investigated. The motivation behind looking at the model parameters, and specifically the intersection of parameters across datasets is as follows. The idea was investigating whether the coefficients associated with the (intersection of the) output parameters of the logistic regression have an estimated value correlated with the accuracy. As such, had these values increased along with the accuracy - and in our case implicitly decreased with time as the highest accuracy belongs to the oldest model - a potential explanation for the phenomenon at hand would be a decrease in the terms' explaining power of the effect on innovation. Taking the intersection of all output parameters, plotting it against their coefficients per dataset and seeing that a trend exists could denote that the parameters lose or gain indicative power, depending on the trend direction. For example, if words such as "business", "new" and "technology" had very high coefficient values in the output of the model ran on dataset 1 as scrapped in 2017, but very low coefficient estimates on the fifth dataset, collected in 2019, one could infer that these words have lost their indicative power with respect to innovation. This is a phenomenon similar to the Google Flu case described in the theory section.

4.2 Results

As a first step in investigating all 4 datasets available, I ran the model using cross-validation (test set is 30 %) in Python. I then looked at the learning rate as the training instances increase. Figure 3 shows the learning behavior of each model, having the number of training samples on the x-axis and the accuracy plotted on the y-axis. The red line represents the training score and the green line represents the validation score. We see that as the training size reaches its maximum, the test and validation lines converge to a point for dataset 1. Importantly, this means that adding new training samples is not likely to increase the model performance for dataset 1, so resources can better be used for building more complex models or investigating different feature selection methods, instead of gathering more data. For the other datasets, the two lines do not yet converge, which means that collecting more training samples could help improve the models.

From analyzing the parameters as explained above one can see several interesting insights. To note immediately is that the loss of informational power does not seem to be the case, since there is no noticeable trend in Figure 4. In Figure 4, the intersection of output parameters is plotted on the x-axis and their corresponding coefficient values on the y-axis, for each of the 4 datasets in a different color. One can firstly see that only 17 parameters are simultaneously occurring in the output of all four models, which is a relatively low number given that the average number of unique words in a corpus is a bit more than 100.000. This fact alone suggests significant parameters thought to be innovative according to dataset 1 have little in common with innovation indicators from the other datasets. In consequence, the indicative power of the parameters changes per in time, which is reflected in the text. Moreover, values for parameters ”bv” and ”us” have an unusually large value, as opposed to the other 3 corresponding coefficients. These peaks prompted further investigate of the word lengths of the output parameters, since these two terms responsible for the peaks are the only ones in the group that have two letters.
The analysis of frequency of short, medium and long words is visualized in different ways in Figures 5 and 6. From Figure 5 one can see the individual histograms per dataset, and in particular the fact that for dataset 1, there seems to be less medium-length words than in the other datasets. Moreover Figure 6 confirms this behavior by reflecting the same information in an overlay density plot. Building on the fact that dataset 1 seems to have an odd behavior in the model output as seen in Figures 4,5 and 6, the subsequent developed strategy was focusing on only two datasets: dataset 1 and dataset 5. Between these two datasets there are several reasons that can cause the accuracy difference. Such reasons are: type of scrapping method, websites present in the corpus, time of collection, text. The choice of only focusing on dataset1 and dataset5 is a means to the goal of holding the effect of the previously mentioned confounders constant, since as seen in Table 1 in the Data section above, both sets have been allegedly collected using Beautiful soup approach in Python. Therefore, only several others remain: the websites present in the 2 corpora, the time of collection, and the text. I moreover control for the individual websites present in each corpora by subsetting only the ones that are common (which results in a sample size of 4189 companies in each set, and each company has it’s representative in the other dataset). This leaves us with two reasons: time of collection and text. Moreover I stress that any differences caused by the time of collection - changes in the websites in real life from 2017 to 2019 respectively - are practically reflected in the text corpus, lexically and/or semantically.
Furthermore, only focusing on the differences in text is reasonable in the aftermath of the results from the parameter analysis above, since they indicate that indeed there might be a difference in text (based on at least word length so far, but maybe based on, or related to, other metrics as well) at hand between dataset 1 and the rest. This choice is also reasonable since the subset of common parameters is unusually small - only 17 in common for all four models, denoting in theory that the terms have lost their informative power in time.
Finally, the problem statement, and the research question for the following parts of this paper has been reduced to: **Is the accuracy of the logistic regression model on dataset 1 (92 %) different from the accuracy of the same model on dataset 5 (62 %) due to or related to the difference in corpora between the two models?**
Section 4.2 describes the strategy employed to detect the differences in text between the two corpora.

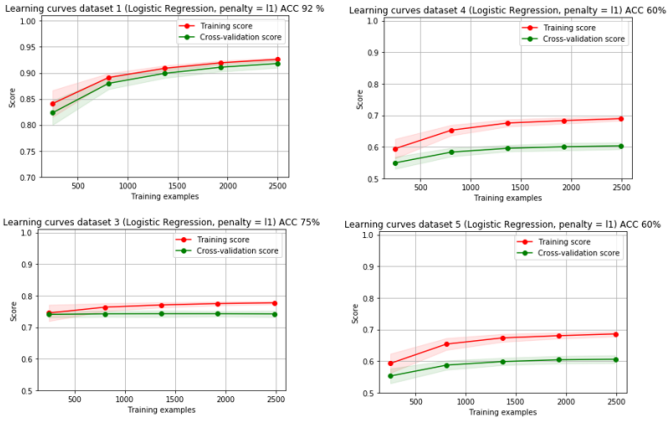

Figure 3: Learning curves for the 4 models.

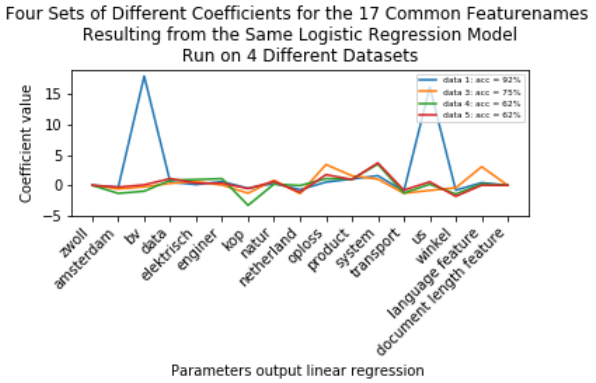

Figure 4: Differences in coefficient estimates between 4 datasets.

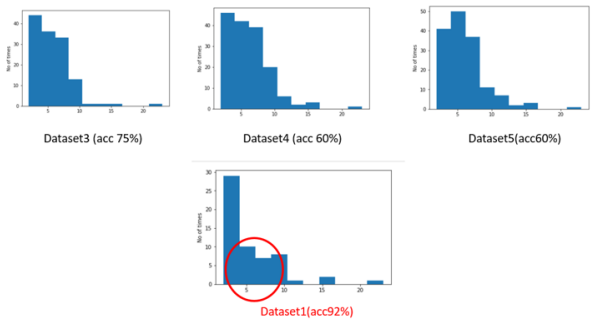

Figure 5: Differences in word length frequency of logistic regression parameters for all 4 datasets.

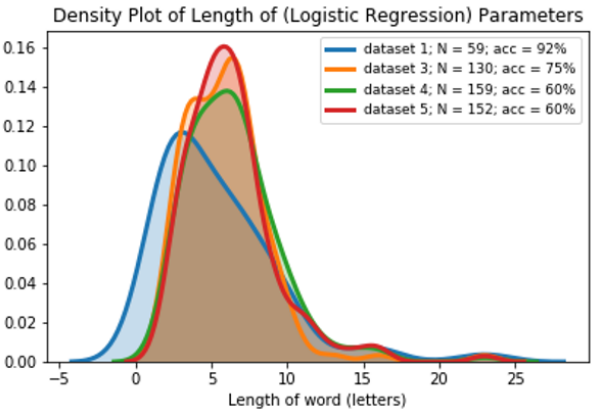

Figure 6: Density plot word length frequency of parameters for all 4 datasets.

5 Analysis II

Because of the different pattern of behavior short, medium and long words seem to have according to Figures 5 and 6 and also importantly because of the very small number of parameters all four datasets have in common, I decided to further investigate the text peculiarities of dataset 1. To hold the other confounders constant, I compare dataset 1 with dataset 5, as they supposedly are created according to the same scrapping procedure. Therefore in this part I aim to investigate differences in text between analogous websites pertaining to datasets 1 and 5 respectively, in order to answer the research question **Is the drop in accuracy from 92 % to 62 % related to the difference in corpora between the two models?** If a difference between texts is found, then we shall conclude that it explains to some degree the differences in model performance.

5.1 Methodology

This section describes the strategy developed to account for possible differences in the text of the two corpora. The analysis is split into two phases: exploration and in depth analysis. During the exploration phase, standard go-to measures were implemented that gave an overview of the general aspect of the data. Next, an in-depth analysis was performed that uses the insights from the first part and also better suited similarity metrics to compare the two datasets. Lastly, the aforementioned advanced similarity metrics have been used to quantify potential differences between types of prediction errors caused by the model on each dataset.

5.1.1 Exploration

A first attempt to see the differences between the two texts was done by performing **K-means clustering** on each individual corpus. K-means is a method of unsupervised learning, that centers data around centroids based on, in this case, euclidean distance. The objective of K-means is the quite simplistic goal to group similar data points together and discover underlying patterns [1]. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset. Based on the elbow method the number of clusters chosen was 2. Elbow method serves as an evaluation for the choice of k, and k is decided visually. The input used for clustering consists of TF-IDF vectors for each corpus, with a term-document matrix shape of (4189,1293) for the first dataset and a shape of (4189,1443) for the fifth, respectively. This difference in size forms the basis of next steps and therefore important to note. A justification for choosing TF-IDF vectors instead of alternatives is that the original model likewise uses these features as input. After the algorithm has centered each the datasets around two clusters, I verified whether these clusters are identical. Based on manual investigation of the documents laying in their set difference, some inferences were made about the potential differences between the two texts. See the appendix for more information on the websites not clustered in an identical manner.

5.1.2 Vector-based similarity measures.

Clustering gives some intuition for group behavior of the two datasets. However, to really grasp the difference between the text representation of each individual website, a better approach is needed. Therefore a method is implemented that can measures pairwise differences between analogous websites, i.e. text_website_i from dataset 1 and text_website_i from dataset 5.

First a bit more details about the **TF-IDF** weights that fill up the term-document matrix that serves as input to the classification model (recall the pipeline). How is TF-IDF computed and what does it mean? In principle, a TF-IDF weight for term t denotes its relevance for document d in a particular corpus.This relevance(importance of term t) increases proportionally to the number of times a word appears in the document and is offset by the frequency of the word in the corpus. [20]. Technically TF-IDF weight is computed using two terms,i.e. by multiplying term frequency (TF) and the inverse document frequency (IDF). To paraphrase an example from [20]. Consider a document containing 100 words in which the word "cat" appears 3 times. The term frequency (TF) for cat is then (3 / 100) = 0.03. Now, assume we have 10 million documents and the word cat appears in one thousand of these. Then, the inverse document frequency (IDF) is calculated as log(10,000,000 / 1,000) = 4. Thus, the TF-IDF weight is the product of these quantities: 0.03 * 4 = 0.12 [20]. Recall that TF-IDF denotes term importance for a document in a corpus. Now, our original empirical assumption stated as motivation for analysis 2 is that the two corpora are identical in text. Should this assumption be indeed satisfied, then also the TF-IDF weights of terms should be identical - which strangely enough they are not, since the matrix corresponding to dataset 1 has shape (4189,1293) and the one corresponding to dataset 5 has shape (4189,1443). The reason why this difference in the length of rows, so 1293 versus 1443, poses a problem is that it prevents more advanced similarity metrics to be applied, i.e. vector based metrics such as cosine similarity, Jensen-Shannon distance, Hellinger distance or Earth Mover Distance (EMD). All of these previously mentioned require as input two vectors of equal lengths.

Therefore another method was originally devised to account for this procedure. My way of addressing this issue was by vectorizing the entire corpora, so the text in dataset 1 together with the text in dataset 5. In this way, TF-IDF produces a matrix twice as large on the row dimension (because each website is accounted for twice). I name this method of vectorization, that incorporates both datasets as a single text corpus, as **global TF-IDF**. Splitting then this large matrix in half on the horizontal axis will result in the two original set of documents, represented with TF-IDF vectorization. Then row 1 from the first half will correspond to row 2 from the second half, where the row in essence represents a website. The advantage of this method is that now comparison *is* possible, as the halved matrices have the same number of columns. Moreover TF-IDF in general affects the overall importance of a term within both a documents and the corpus. It is implied that, should two websites be identical in terms of text, then the individual TF-IDF vectorization is identical to the global TF-IDF vectorization, which justified this vectorization method. An additional reason for choosing to transform terms into vectors using TF-IDF (and global TF-IDF) is that this is the way the original model was created in 2017, and also part of the accuracy decrease problem. Nevertheless, better vectorization methods do exist for the current issue. Such an example is **count vectorization**: arguably the most intuitive way of preprocessing words in NLP, simply each unique word in the corpus represents one column, and each element in the matrix denotes how many times the respective word occurs in the row/document it is in. The advantage of the count method for the present case is that it allows for a more clear overview of all words, since all unique words are present; especially even if there is one word present in one document in dataset 1 but absent in it's twin corresponding document from the other dataset, it will still be denoted with a 0 in the former, and with a 1 in the latter. This small but important fact allows comparison using vector-based metrics, since it the property of identical rows/vectors is satisfied.
After the globally vectorized matrix G has been split into equal, equivalent halves H1 and H2, all four vector-based similarity metrics (cosine, EMD, Jensen-Shannon and Hellinger) have been computed for each pair of analogous websites text. So if index 123 in matrix H1 corresponds to website *i* we first of all know for sure that index 123 in matrix H2 corresponds to the same website i, which is a particularly nice and useful property. Furthermore we are trying to prove that the text in these two analogous website "versions"is intrinsically different in matrix H1 versus H2, as per the research question.

5.1.3 Difference between prediction error groups.

The procedure mentioned earlier, of global vectorization, has been performed twice, one using TF-IDF vectorization and once using count vectorization. The reason for using count vectorization as well is justified by the fact that it makes most sense in relation to the Jensen-Shannon Distance, which compares distribution. Using the counts, each word exists in the distribution - and if it only exists for example in dataset 1 but not in dataset 5, the count representation will be 0 for the former and 1 for the latter. This view is more comprehensible than TF-IDF vectors. On the other hand, TF-IDF vectorization was used , despite it being less obvious as a comparison method, because the original model I have been provided with uses TF-IDF vectors as well in the input for the logistic regression, and is therefore consistent with the original approach.
The next step, after computing differences between equivalent websites in the two datasets has been to append these results to the prediction error dataset. To explain,

prediction errors in this case are split into four groups: websites that are in reality innovative but according to dataset 5 are not innovative (False Negative 5), websites that are in reality not innovative but according to dataset 5 are indeed innovative (False Positives 5), websites that are in reality innovative but according to dataset 1 are not innovative (False Negative 1) and websites that are in reality not innovative but according to dataset 1 are indeed innovative (False Positives 1). Interesting to note is that the latter group is not present in the current case. This leaves further analysis with the other three: False Negatives 5, False Positives 5 and False Negatives 1.
In order to detect whether there might be a significant difference between these 3 groups in terms of either one of the four implemented similarity metrics, some statistical tests were performed for unbalanced data. Note that the sample sizes are not equal, i.e. data is unbalanced (N$FN5$ = 169, NFP5 = 225 N$FN1$ = 37)

5.2 Results

First the results from the overview will be presented, and then the results from the in-depth analysis. Interesting to note here as well is the fact that looking at the document length variable, 94.3 % of all websites have different lengths. Although the time of collection is different so it might be reasonable, this is still worth mentioning.

5.2.1 Exploration

From the **K-means clustering exploration** , the result is that each dataset is divided into two clusters, of comparable sizes. However the corresponding clusters do not match entirely. There are 10 documents, i.e. websites that belong into clusters 1 and 2 in dataset 1 are centered around the opposite clusters in the second dataset. See Figure 7. All of these 10 documents were inspected manually and the corresponding dataframes are presented in Appendix A. Looking at these document representations, some insights were noticed.
First, most websites have shorter document lengths in dataset 1. A notable example is website with index 37751336, for which the document length is 50 in the first dataset, and 164 in the fifth. By looking at the text and using the internet, this website has been identified to belong to Royal Fassin, a company that commercializes in candy (fruit gum and liquorice products).
A second remark seen from investigating the set difference from the clusters is that, in this subset of 10 websites, at least 4 accounts of concatenated terms have been observed. In a concrete example from the same website, Royal Fassin provided text in dataset 1 has terms such as ”fassinproductenkwaliteitmediacontact” , ”visievacaturesgoed”, ”doelenproductenbusinessov” - terms which do not appear in the same website’s version of stemmed text in dataset 5. See Appendix D for the whole text corresponding to Royal Fassin.
Third, the language detection is not informative when it comes to English, since German text is also identified as English. For an example see in Appendix C company indexed 12305162. If some may associate a company’s promotion language as English to be more open and innovative (which would be reasonable to assume), perhaps less people would infer the same for websites that have their websites in German. In fact, the relationship might even be switched: as Germans are a nation who rarely speak a second language, promoting a company in German might mean the company has a closed strategy, perhaps only targeted at the German market. This company might also be innovative, but for other reasons than a website in English would.

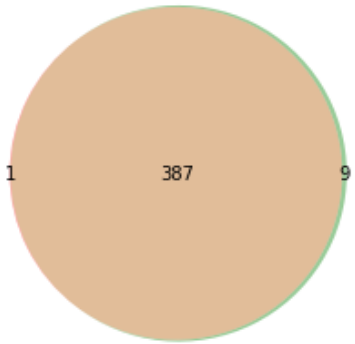

Figure 7: Intersection of K-means clustering (k=2) for dataset 1 and dataset 5. There are 10 websites that do not fall into same clusters.

5.2.2 Vector-based similarity measures.

Each pair of twin websites has attached to it a value according to Jensen-Shannon distance, where 0 means the two distributions are identical. This gives a total of 4189 Jensen-Shannon values, ranging from 0 to 1. Furthermore as explained in the methodology, there is a subset of prediction errors of size 431.
Based on the Central Limit Theorem (CLT) the mean of the average distance is normally distributed. I estimate the standard error of the sampling distribution of the average distance through bootstrapping. Furthermore the 99% confidence intervals for the average distance are computed. By observing the confidence intervals from Table 2, we can see that each interval is very narrow. I stress that the focus of this method is on the count vectorization and Jensen-Shannon similarity metric. The others were also calculated for reasons of consistency, but the most meaningful and the most important distance metric is Jensen Shannon, since it compares distributions and is bounded as opposed to Hellinger which is not. Also counts are more objective than TF-IDF as explained earlier; TF-IDF was also implemented for each metric because TF-IDF was the initial approach for preprocessing of the data as explained in the Pipeline section.
We have as an empirical assumption, i.e. the null hypothe-

sis in this case, that each pair of twin websites should have identical text contents, i.e. a Jensen-Shannon metric of 0. The rejection of the null hypothesis can be done if the value 0 is not in the 99 % confidence intervals resulting from the bootstrapping procedure. Since we see that in Table 2 the confidence intervals of the Jensen-Shannon estimated average pertaining to the count vectorization method is (0.2353, 0.2530) and (0.2164, 0.2317) for TF-IDF respectively we conclude that the null is rejected and indeed the two pairwise websites are different from each other (Jensen -Shannon statistically different than 0, at 99 % confidence). A similar statistical inference can be done for the other metrics in Table , with the only remark that for the cosine similarity metric, 1 denotes identical documents, so translating the empirical assumption that the two websites are identical is done with stating the null hypothesis "Cosine (Website i (dataset 1), Website i(dataset 5)) = 1. " and the alternative hypothesis being "Cosine (Website i (dataset 1), Website i(dataset 5)) smaller than 1. "
We nevertheless see that in each case, according to the Central Limit Theorem, the documents or websites pertaining to the predictions error group are always containing different underlying texts, assessment made at 99 % confidence level.

Vect.	Similarity Metric	Average Estimate	Confidence Intervals (99 %)
Count	Cosine	0.747	(0.7378, 0.7564)
	EMD	0.000710	(0.0007, 0.0008)
	Jensen-Shannon	0.244	(0.2353, 0.2530)
	Hellinger	9.15	(8.910, 9.391)
TF-IDF	Cosine	0.709	(0.6977, 0.7193)
	EMD	0.000706	(0.0007, 0.0007)
	Jensen-Shannon	0.224	(0.2164, 0.2317)
	Hellinger	1.37	(1.341, 1.402)

Table 2: Estimates of average similarity metrics on count vectorized input (bootstrapped, N=5000 simulations).

5.2.3 Difference between prediction error groups.

Testing for the most dissimilar group amongst the prediction errors (false negatives and false positives for dataset 5 and false negatives for dataset 1), several tests were tried and performed.
First, since ANOVA could not be performed due to the fact that its main assumption of normality of data does not hold, as seen in the Shapiro-Wilk results and p-values in the tables from Appendix F. Note that the only p-value that is bigger than 0.05 corresponds to the Hellinger Distance on TF-IDF vectorization, for the predictions error subset False Negatives from dataset 5 (Appendix F). However, this is only one of three factors (the other two being the other two types of prediction errors available, i.e. false positives and negatives respectively corresponding to dataset 1). Testing the whole set of Hellinger values for the entire set of prediction errors gives a p-value of 8.582e-09, which is not greater than 0.05 and therefore fails reject the null hypothesis according to which the data is normally distributed. Since the normality assumption does not hold for any of the metrics, it is not appropriate to run parametric tests for differences between groups such as ANOVA.
In consequence, a non-parametric test, which does not assume normality was performed. Kruskall-Wallis Rank Sum is such a test (Kruskall-Wallis results in Table 3) . The null hypothesis assumes that the samples (groups) are from identical populations, and the alternative hypothesis is that at least one of the groups stems from a different population than the others. Since none of the p-values (Table 3) are less than 0.05, we can not reject the null hypothesis, so all prediction error groups (FN1, FP1, FN5) are coming from the same distribution.

Similarity Metric	Kruskal-Wallis chi-squared	p-value
Cosine Similarity Metric (tf-idf)	0.28806	0.8659
Earth Mover Distance (tf-idf)	1.1206	0.571
Jensen-Shannon Distance(tf-idf)	0.3085	0.8571
Hellinger Distance (tf-idf)	0.16722	0.9198
Cosine Similarity Metric (count)	0.43805	0.8033
Earth Mover Distance (count)	0.08207	0.9598
Jensen-Shannon Distance(count)	0.4067	0.816
Hellinger Distance (count)	0.36346	0.8338

Table 3: Kruskal-Wallis for all similarity metrics for 3 factors, i.e. 2 degrees of freedom.

6 Discussion

From analyzing all 4 datasets the conclusion was that the model ran on the first has a different cross-validation behavior than the other three. Since the training and the test line do converge, a recommendation for the product owner would be investing in a more complex model rather than in additional data. Furthermore from the clustering approach we have seen that dataset 1 scrapping output is unusual since it contains concatenated words that clearly do not exist either in the scrapped output of dataset 5, nor in a real life setting, such as "fassinproductenkwaliteitmediacontact", "visievacaturesgoed", "doelenproductenbusinessov" (see Appendix D). From this we can recommend a closer attention to the scrapping method, and employing human-in-the-loop methods such as manual checks that detect such anomalies. Otherwise,

saving the websites and replicating the scrapping methods and then checking also using similarity metrics such as those presented here can help reveal differences and anomalies between scrapping outputs. Next, also from the second Analysis we have seen that on the prediction error subset, representation of a website in dataset 1 is significantly different that it's analogous representation in dataset 5. This prompts the conclusion that indeed the differences in accuracy of the two models can be said to be related to the intrinsic differences in text caused by scrapping. A closer look also brings TF-IDF vectorization method to a disadvantage when it comes to modelling text and specifically building predictive models from text that is likely to have high volatility, i.e. change with time. For such types of text, it is better to use vectorizations and embeddings that are less dependent on the initial textual structure - the naive example that would work as a good alternative here is count vectorization, because all words are by definition accounted for, even if just with a value of 0. More sophisticated examples are deep learning based methods, part of the knowledge based vectorizations, such as word2vec or glove. Lastly from the second analysis we conclude that the three prediction groups are not statistically different from one another: based on the methods employed here, we can not say for sure which of these groups (false negatives 1, false positives 1 or false negatives 5) is most problematic, i.e. has similarity values that point to different texts more than the other. With more data, the analysis can be replicated and the larger group sizes could give more information about potential differences.

A strength of this research paper is that it gives yet another example of big data related issues, based on a real life application. This will further inform both academics and industry data scientist of pitfalls they are likely to fall into. As a weakness, it is useful to note that the original logistic regression model exclusively receives big data as input (text from websites). It would have been interesting to include some factual information about innovation in Dutch companies as explanatory variables as well. These added predictors would attenuate big data hubris (the implicit assumption that big data are a substitute for, rather than a supplement to traditional data collection and analysis [15]). In the new envisioned model, empirically based metrics (new products delivered by company i in the past 2 years, budget for innovation, is there a department for innovation, international employees, etc.) designed in collaboration with experts in innovation fields would be included alongside a metric for NLP. Then the NLP metric could be tested and checked for significance as usual. Ideally, using this approach would show that neither the empirical-only predictors, nor the big data-only predictors have the best performance, but a combination of both as seen in the Google Flu case [15] [10].

In the introduction I introduced the model pipeline that was used for classification of innovation in Dutch companies based on the text on their websites. Now, after I have presented the analysis results it could be useful to point where the problem might have occurred in this pipeline.

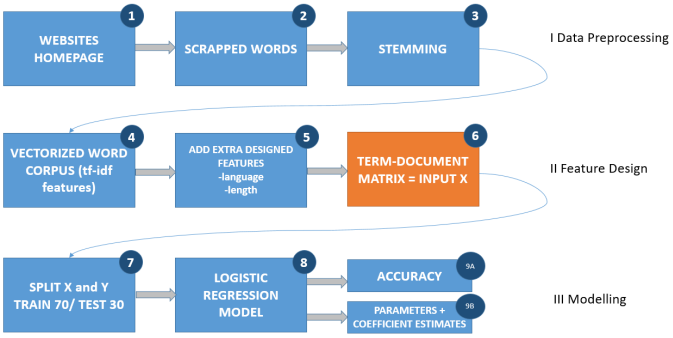

Figure 8: Risk zones in project pipeline. Analysis II is motivated on the issue of differently sized term-document matrices for datasets 1 and 5. (block 6 in the figure)

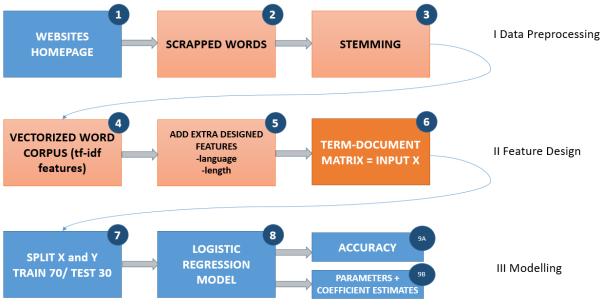

Figure 9: Risk zones in project pipeline, extended. Although the problem can be traced back to the differences in the term-document matrix (6) - Analysis II essentially starts from the fact that TF-IDF vectorizations of the different datasets have different sizes - this difference can in turn be understood to have its basis in the intrinsic differences in scrapped words (2).

First, as the introduction and motivation of applying vectorizations to the global dataset, so dataset 1 concatenated with 5, instead of doing this step to each individually, we remark that Term-Document Matrix which acts as the input X to the original logistic regression model, is where the difference in accuracy can be explained. Moreover, by analyzing this term-document matrix using four different similarity metrics we can go even further into the pipeline and be sure that the difference occuring in the matrix lays in the text difference, by the global vectorizations approaches. This is an intriguing result. After all, the scrapping method using Pythons' Beautiful Soup applied to dataset 1 originary seems to be different than the scrapping method from dataset5 (See again Table 1). There is however a silver lining. If indeed that is the case, this issue is easily avoidable - for example one could replicate the scraping on a sample of websites and then , perform the global vectorization as described in methodology of analysis 2. If the similarity between the two is high one

can tell that scraping is consistent and data collection has been performed well.
It is worth discussing step 4 from the pipeline in more detail. I stress again that for comparing text, count vectorization is the best method to transform words into numbers. This is because every word is accounted for, and words that do not exist in one text but do in the other will simply score a 0 according to this bag-of-words method.
Another limitation of this study is its inability to asses potential causes or risk zones that stem from part 1 of the pipeline, i.e. the real websites homepages. Firstly, the data that served as object of my analysis did not include the original datasets, because for dataset 1 they were not saved. Secondly, the lack of repeated measurements of the same websites across time (time-series data) automatically diminishes the possibility to evaluate the long-term generalizability of the model. A recommendation is to gather such data and see how the performance of the classification model evolves with time.
And last but not least, a small remark about part 8 of the pipeline, i.e. the model. Although big data is at it's peak in terms of prevalence, importance and relevance in daily life, business and academia, there still seems to be, amongst business decision-makers but professional researchers alike, an unfortunate preference for models that are more interpretable than they are accurate [3]. The dilemma posed in the introduction, that represented nature as a black box is then translated - in the algorithmic approach - as just another black box (random forests, neural networks, etc.) I quote from [3]: My biostatistician friends tell me, "Doctors can interpret logistic regression." There is no way they can interpret a black box containing fifty trees hooked together. In a choice between accuracy and interpretability, they'll go for interpretability. It is uncanny how many such almost identical remarks are en vogue even 18 years after Breiman's paper. However, it has been proven that in certain cases this black box gives better cross-validation results than logistic regression [3]. Moreover, potentially better performance of a more complex model - at least for dataset 1- is sustained by the learning curves results from Analysis I (see Figure 3). Since the goal here is to produce a good estimate - for example for small innovative companies - the relevance of some coefficients of a logistic regression should be lower than the relevance of accuracy, i.e. getting the right number of such small innovative companies.
To conclude: big data is evolutionary and can provide innovative, real-time, and more granular insight for economic and financial analysis. Its potential can not be overstated, however neither are its challenges. These challenges (multi-dimensionality of data, spurious correlations, abundance of false positives, etc.) have been researched and discussed by many [10] [15] [6] and reviewed in this paper in the theory section as well. Still, beyond differences that big data has from traditional data in terms of statistical inference on one hand and storage on the other [6], even people who are called data scientists can have different perspectives about the same problem. Here for example the pipeline is long and few people have expert knowledge in each step along the way. This essentially should not be a problem, provided communication and replication is implemented in a systematic way, and provided - at least for NLP projects such as the ones discussed - the risk zones I have highlighted are given additional attention.

7 Acknowledgments

The author of this paper has been an intern for the Centre for Big Data Statistics of Centraal Bureau voor de Statistiek (CBS) while working on this project. Special thanks to my friend Dakai Wei, fellow student in the Data Science Bachelor and fellow intern at CBS, for his constant stream of solicited and unsolicited advice, both on my methods and on structuring this paper. I would also like to thank my project supervisor Piet Daas, for all the suggestions and support but also for his forever inspiring and optimistic vision on the marriage of big data and traditional statistics.

References

[1] Andrey Bu. Machine learning model: python sklearn and keras. https://www.education-ecosystem.com/andreybu/REaxr-machine-learning-model-python-sklearn-kera/oPGdP-machine-learning-model-python-sklearn-kera/.

[2] Mohamed Arouri, Amal Aouadi, Philippe Foulquier, Frédéric Teulon, et al. Can information demand help to predict stock market liquidity? google it! Technical report, 2013.

[3] Leo Breiman et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

[4] John Dunne and S MacFeely. Big data coming soon..... to an nsi near you. In *World Statistics Congress, 25th-30th, Hong Kong. http://www. statistics. gov. hk/wsc/STS018-P3-A. pdf (last accessed 1st April 2015)*. Citeseer, 2013.

[5] Jianqing Fan and Yingying Fan. High dimensional classification using features annealed independence rules. *Annals of statistics*, 36(6):2605, 2008.

[6] Jianqing Fan, Fang Han, and Han Liu. Challenges of big data analysis. *National science review*, 1(2):293–314, 2014.

[7] Denisa Florescu, Martin Karlberg, Fernando Reis, P Rey Del Castillo, Michail Skaliotis, and Albrecht Wirthmann. Will big datatransform official statistics. In *European Conference on the QualityofOfficial Statistics. Vienna, Austria*, pages 2–5, 2014.

[8] Vanessa Frias-Martinez, Cristina Soguero-Ruiz, Enrique Frias-Martinez, and Malvina Josephidou. Forecasting socioeconomic trends with cell phone records. In *Proceedings of the 3rd ACM Symposium on Computing for Development*, page 15. ACM, 2013.

[9] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012, 2009.

[10] Sharad Goel, Jake M Hofman, Sébastien Lahaie, David M Pennock, and Duncan J Watts. Predicting consumer behavior with web search. *Proceedings of the National academy of sciences*, 107(41):17486–17490, 2010.

[11] Wael H Gomaa and Aly A Fahmy. A survey of text similarity approaches. *International Journal of Computer Applications*, 68(13):13–18, 2013.

[12] Peter Hall, Yvonne Pittelkow, and Malay Ghosh. Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):159–173, 2008.

[13] Cornelia Hammer, Ms Diane C Kostroch, and Mr Gabriel Quiros. *Big Data: Potential, Challenges and Statistical Implications*. International Monetary Fund, 2017.

[14] Yigitcan Karabulut. Can facebook predict stock market activity? In *AFA 2013 San Diego Meetings Paper*, 2013.

[15] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of google flu: traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.

[16] Rada Mihalcea, Courtney Corley, Carlo Strapparava, et al. Corpus-based and knowledge-based measures of text semantic similarity. In *Aaai*, volume 6, pages 775–780, 2006.

[17] Royal Fassin Website. Royal fassin. https://www.royalfassin.com/.

[18] Noam Shoval and Rein Ahas. The use of tracking technologies in tourism research: the first decade. *Tourism Geographies*, 18(5):587–606, 2016.

[19] Thomas H. Davenport. Data scientist: The sexiest job of the 21st century. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century.

[20] Unknown. Tf-idf :: A single-page tutorial - information retrieval and text mining. http://www.tfidf.com/.

[10] [13] [15] [1] [17] [7] [19] [11] [16] [2] [14] [3] [9] [6] [5] [12] [20] [4] [8] [18]

8 Appendix A.

Feature Code	Survey Data	Administrative Data	Big data
F1	Statistical products specified ex-ante	Statistical products specified ex-post	Statistical products specified ex-post
F2	Designed for statistical purposes	Designed for other purposes	Organic (not designed) or designed for other purposes
F3	Lower potential for by-products	Higher potential for by-products	Higher potential for by-products
F4	Classical statistical methods available	Classical statistical methods available, usually depending on the specific data	Classical statistical methods not always useful
F5	Structured	A certain level of data structure, depending on the objective of data collection	A certain level of data structure, depending on the source of information
F6	Weaker comparability between countries	Weaker comparability between countries	Potentially greater comparability between countries
F7	Representativeness and coverage known by design	Representativeness and coverage often known	Representativeness and coverage difficult to asses
F8	Not biased	Possibly biased	Unknown and possibly biased
F9	Typical types of errors (sampling and non-sampling errors)	Typical types of errors(non-sampling errors e.g. missing data, reporting errors and outliers)	Both typical errors (e.g. missing data, reporting errors and outliers) although possibly less frequently occurring, and new types of errors
F10	Persistent	Possibly less persistent	Less persistent
F11	Manageable volume	Manageable volume	Huge volume
F12	Slower	Potentially faster	Potentially much faster
F13	Expensive	Inexpensive	Potentially inexpensive
F14	High burden	No incremental burden	No incremental burden

Table 4: Main features of survey, administrative and big data; adapted from [7]

9 Appendix B.

Data Source +	Data Type	Statistical Domains	Additional Statistical Domains*
Social Networks	1100. Social Networks: Facebook, Twitter, LinkedIn	National accounts	Sentiment indices (investor, consumer)
	1200. Blogs and comments	External sector statistics	Social statistics
	1600. Internet searches on search engines (Google)	Financial statistics	Labor statistics
	1700. Mobile data content: text messages, Call Detail Record, Data Detail Record, Location update, Radio coverage updates Online news	Price statistics; Government finance statistics (public debt statistics)	Migration statistics
Traditional Business Systems	**Data produced by public agencies**	Government finance statistics;	Sustainable Development Goals indicators
	Administrative data	National accounts; Price statistics; External sector statistics	
	Data produced by businesses	National accounts	Social Statistics
	2210. Commercial transactions	Price statistics	Business registers
	2220. Banking/stock records	External sector statistics	Employment statistics
	2230. E-commerce	Financial statistics	Transport statistics
	2240. Credit cards		Sustainable Development Goals indicators
	Business websites Scanner data		
Internet of Things(machine-generated data	**Data from sensors**	National accounts	Traffic/transport statistics
	311. Fixed sensors	Satellite accounts	Energy statistics
	3111. Home automation	External sector statistics	Land use statistics
	3112. Weather/pollution sensors	Government finance statistics	Agricultural statistics
	3113. Traffic sensors/webcam	Price statistics	Environment statistics
	3114. Scientific sensors		Transport and emission statistics
	312. Mobile sensors (tracking)		Air emission statistics
	3121. Mobile phone location		Sustainable Development Goals indicators
	3122. Cars 3123. Satellite images		

10 Appendix C.

The documents (N=10) that are not clustered in the same fashion by K-means (k=2) on vectorized input are presented here. I show how these websites are represented differently in dataset 1 and dataset 5.

BEID	lang	Innov	text	doc length
16821327	english	1	intralox conveyor belt equip servic intralox s...	473.0
23731958	english	1	dutch farmholiday stay farm bed breakfast line...	218.0
37751336	english	1	fascini world fascini homeworld fascininlenlog...	50.0
68554230	english	1	swt wast technolog jump content homeoverproduc...	127.0
10461647	english	0	est lauder nederland welkom registreren schrij...	335.0
12305162	english	0	ihr innovationspartn ndelt sein kompetenzen nu...	168.0
16521447	english	0	superuni superuni contact toggl navig superuni...	123.0
18227686	english	0	telecomspecialist van headset tot telefooncent...	765.0
27072169	english	0	home page javascript lijkt uitgeschakeld zijn ...	576.0
56983328	english	0	konica minolta netherland skip content deze we...	578.0

Table 5: Documents not clustered identically by K-means. Representation in dataset 1.

BEID	lang	Innov	text	doc length
16821327	english	1	intralox conveyor belt equip servic intralox s...	585.0
23731958	english	1	dutch farmholiday stay farm bed breakfast line...	221.0
37751336	english	1	welcom royal fassin welcom royal fassin stay a...	164.0
68554230	english	1	swt nl en een divisi van sma menu swt referent...	572.0
10461647	english	0	este lauder beauti product skin care makeup me...	79.0
12305162	english	0	nuscienc group royal agrifirm group agrifirm c...	369.0
16521447	english	0	superuni superuni contact toggl navig superuni...	129.0
18227686	english	0	telecomspecialist van headset tot telefooncent...	556.0
27072169	english	0	bombeeck digit groothandel digital tv ontvangs...	874.0
56983328	english	0	konica minolta netherland skip content deze we...	552.0

Table 6: Documents not clustered identically by K-means. Representation in dataset 5.

11 Appendix D.

Text persisting of 50 terms corresponding to Royal Fassin (BEID = 37751336) in dataset 1:
”fascini world fascini homeworld fascininlenlogin homeov fassinproductenkwaliteitmediacontact world fascini world fascini kwaliteit world fascini world fascini fassin meer dan jaarfascinati royal fassin fassin vacatur vacatur producten producten world fascini medewerkerslogin client login producten producten world fasciniov fassin jaar fascinatiemissi visievacaturesgoed doelenproductenbusinessov fassinproductenroy fassin box aa heerenbergulenpasweg gb heerenberg nederlandphon fax”

Text persisting of 167 terms corresponding to Royal Fassin (BEID = 37751336) in dataset 5:
”welcom royal fassin welcom royal fassin stay amaz passion get fascin royal fassin dutch compani base heerenberg near german border royal signatur award us behalf queen netherland th anniversari us repres strong sens long term vision confid reliabl social respons engag royalti therefor alway reson proud independ famili busi royal fassin special product extrud fruit gum liquoric product modern techniqu dedic peopl highest qualiti standard make happen truli passion do passion goe back year joseph langenberg xaver fassin found compani journey start spark never end treat new websit come soon current temporari websit make fassin experi thrill be develop brand new websit thank patienc cours keep post meantim ism trade fair readi sweet talk meet us end januari ism tradefair januari th januari th cologn germani stand hall download latest sweetest brochur catalogu download view onlin get fascin download view onlin valuabl brand contact us address royal fassin box aa heerenberg ulenpasweg gb heerenberg netherland phone email info royalfassin com view privaci statement cooki”

12 Appendix E.

Count Vectorization	**Cosine Similarity**	**EMD**	**Jensen-Shannon Distance**	**Hellinger Distance**
False Negatives 5	0.7610	0.00063	0.2355	9.221
False Positives 5	0.7501	0.00074	0.2473	9.017
False Negatives 1	0.7673	0.00072	0.2196	8.7103

Table 7: Average similarity metrics on count vectorized input, for each of the 4 similarity metrics.

TF-IDF Vectorization	**Cosine Similarity**	**EMD**	**Jensen-Shannon Distance**	**Hellinger Distance**
False Negatives 5	0.7185	0.00066	0.2192	1.3347
False Positives 5	0.7099	0.000732	0.2233	1.3477
False Negatives 1	0.7307	0.000741	0.2002	1.2938

Table 8: Average similarity metrics on TF-IDF vectorized input, for each of the 4 similarity metrics.

13 Appendix F.

False Negatives 1	Shapiro-Wilk test statistic	p-value
Cosine Similarity Metric (tf-idf)	0.86674	4.379e-11
Earth Mover Distance (tf-idf)	0.86411	3.254e-11
Jensen-Shannon Distance(tf-idf)	0.89152	8.842e-10
Hellinger Distance (tf-idf)	0.96094	0.0001133
Cosine Similarity Metric (count)	0.8725	8.505e-11
Earth Mover Distance (count)	0.53248	2.2e-16
Jensen-Shannon Distance(count)	0.88427	3.519e-10
Hellinger Distance (count)	0.9307	2.935e-07

Table 9: Normality test for similarity values corresponding to prediction error False Negatives, Dataset 1.

False Positives 1	Shapiro-Wilk test statistic	p-value
Cosine Similarity Metric (tf-idf)	0.87906	2.055e-12
Earth Mover Distance (tf-idf)	0.84373	2.607e-14
Jensen-Shannon Distance(tf-idf)	0.90389	7.703e-11
Hellinger Distance (tf-idf)	0.96081	7.623e-06
Cosine Similarity Metric (count)	0.8835	3.777e-12
Earth Mover Distance (count)	0.3621	2.2e-16
Jensen-Shannon Distance(count)	0.85938	1.642e-13
Hellinger Distance (count)	0.92872	5.664e-09

Table 10: Normality test for similarity values corresponding to prediction error False Positives, Dataset 1.

False Negatives 5	Shapiro-Wilk test statistic	p-value
Cosine Similarity Metric (tf-idf)	0.80632	1.751e-05
Earth Mover Distance (tf-idf)	0.9004	0.003024
Jensen-Shannon Distance(tf-idf)	0.85423	0.0001979
Hellinger Distance (tf-idf)	0.96081	0.4863
Cosine Similarity Metric (count)	0.82596	4.536e-05
Earth Mover Distance (count)	0.6513	3.943e-08
Jensen-Shannon Distance(count)	0.85347	0.0001898
Hellinger Distance (count)	0.92212	0.01288

Table 11: Normality test for similarity values corresponding to prediction error False Negatives, Dataset 5.