Discussion Paper

Classifying businesses by economic activity using web-based text mining

2017 | 18

Maarten Roelands,
Arnout van Delden
Dick Windmeijer

Content

1. Introduction 4

2. Approach 7

3. Results 25

4. Discussion 37

5. Conclusions 41

6. Appendix 42

7. References 48

Acknowledgements 51

Summary

For National Statistical Institutes determining the economic activity is an ambiguous and therefore difficult classification task. With the growth of the amount of available data on the web and big data techniques, automatic classification of economic activity to enrich the now available classifications is a promising technique. The purpose of the present study is to evaluate the suitability of text mining techniques to classify economic activity based on texts extracted from business web sites. We used a case study that classifies businesses into so-called top-sector categories. This classification consists of 9 main categories and 29 subcategories. Businesses that do not belong to one of those top-sector (sub)categories were appointed to an additional category called “other”.

We evaluated a number of methodical aspects: the use of a multi-label versus a single-label prediction, the use of a knowledge-based versus an automatic feature selection, the performance of different classifiers. We also compared the performance of text mining for different subpopulations: one-man businesses versus larger businesses and classification derived from NACE-codes versus a classification based on trade organisation membership.

Starting point of our study was a population frame in which all enterprises were appointed to a top-sector class. For most of the enterprises this top-sector code was directly derived from its NACE code. The other enterprises were found on membership lists of trade organisation, and their top-sector code was assigned manually.
We used this population frame to draw a sample stratified by the sub-sector code, with a net sample size of 1918 enterprises. This sample was split into a training-validation set and a test set. We then predicted the top-sector codes and the sub-sector codes by using supervised machine learning methods.

In our case study, the feature selection based on the pooled results of the knowledge-based and an automatic feature selection yielded the best results. A Naïve Bayes classifier performed better than other classifiers that were tested: *k*-nearest neighbour, random forest, support vector machine and logistic regression. We obtained an accuracy of 51% for our best performing method at top-sector level while that for sub-sector level was much smaller. In the discussion, we present several ideas to improve the performance.

Keywords

Text Classification, Economic activity, Supervised Machine Learning, Naive Bayes, *k*-nearest neighbour, Random Forest, Support Vector Machine, Logistic Regression, Feature Selection,

1. Introduction

National Statistical Institutes continuously stand for the challenge to produce statistics that are rich in information content and reliable for society. Due to the raised information demand by society, National Statistical Institutes search for ways to enrich the available data to improve available statistics. For instance, big data techniques have been used for retrieval of price index information (Griffioen et al., 2014; Struijs et al., 2014; Reimsbach-Kounatze, 2015). Big data sources, coming from both the world wide web and from sensors in electronic devices, deliver interesting insightful information that can be used on its own or in combination with already existing (traditional) primary data sources (Buelens et al., 2014). Big data sources might be used to partially substitute the current data sources (Tam and Clarke, 2015) or, in combination with other data sources they may be used for data validation (Cheung, 2012, Hackl, 2016).

This is also the case for business statistics. In today's situation National Statistical Institutes need to combine different sources and/or held cost intensive surveys to obtain data about business activities.

A core characteristic of enterprises is their economic activity, according to the NACE rev 2 classification. Output of business statistics is are often grouped by industries which follow from the NACE codes. Since a number of years, National Statistical Institutes and third parties are interested in additional classifications to characterise groups of businesses, such as businesses with 'corporate social responsibility', 'family businesses' and 'innovative businesses'.

Website information may be a useful source to derive such alternative business classifications. Website information is also a potential source to improve the quality of the currently appointed NACE codes. The current NACE codes are often based on administrative data, for instance chamber of commerce data where businesses register themselves when they start their business. The NACE code is often not up to date, since businesses may gradually change their economic activities but they seldom report this change to the chamber of commerce. It is also complex to determine the correct economic activity, for instance because a business may have multiple activities. The quality of the current NACE codes may be improved when different sources of information on the NACE code are combined.

Text mining techniques may be used to automatically derive economic activity from web site information. Research on the use of text mining to automatically derive economic activity and occupation from survey answers had been studied by, for instance, Gweon et al. (2017), Jung et al. (2008) and Thompson et al. (2012). However, to the best of our knowledge, little research has been done so far on the suitability of text mining to automatically derive economic activity from website information. We believe that at least part of the business websites contain information on their economic activity since businesses need to profile themselves

through their websites and more and more economic activities are conducted online. Since most online information is in textual format text mining applications can be used to extract those information.

The long-term aim of this study is to develop a method to automatically derive economic activity from information on business websites. From a societal point of this long-term aim is worth addressing, since it is a means to enrich the available information at a relatively low cost (Hand, 1998; Cheung, 2012; Daas et al., 2015; Struijs et al., 2014; Hassani et al., 2014). Before these benefits can be achieved, we need to find out if this automatic classification can be done with sufficient performance.

From a scientific point of view this is also interesting, because there are only a limited number of applications of big data methods in official statistics (Daas et al., 2015). Experiences from text-mining applications have shown that their success is very data specific (Hearst, 2003; Daas, 2012; Aphinyanaphongs et al., 2014; El-Halees, 2015; Tam and Clarke, 2015).

There are many studies on automatic classification of industry and occupation coding, see for instance Chen et al. (1993), Gweon et al. (2017), Jung et al. (2008), Tarnow-Mordi (2017), Thompson et al. (2012) and references therein. To the best of our knowledge, these studies are limited to the situation where answers are given to open questions in survey sampling by persons or representatives of businesses. In these studies, a number of machine learning methods are used, such as support vector machines (Tarnow-Mordi, 2017), *k*-nearest neighbour (Gweon et al., 2017), maximum entropy models (Jung et al., 2008) and logistic regression (Thompson et al., 2012).

In the present study, we explore the potential of text mining methods based on website information. We need to deal with four complicating factors. Firstly, in contrast to the answers to open-ended questions, websites texts are not designed to describe economic activity. Moreover, enterprises may have multiple economic activities, but we are only interested, at this stage, in the *main* economic activity. The challenge here is to distinguish “signal from noise”. Secondly, websites vary in the amount of words they contain, in their language and in their structure. Thirdly, the correct economic activity of an enterprise is hard to determine since enterprises may have a mixture of economic activities, the classes of the classification of activities are not always completely disjoint and some classes have a narrow definition while that of others is rather wide. Fourthly, it is hard to obtain an error-free learning set of sufficient size, since it is time-consuming to (manually) determine the correct economic activity of an enterprise.

The objective of the current paper is to evaluate the suitability of text mining techniques to automatically classify enterprises to a standard classification of economic activity. As an example of a standard classification, this paper will use the classification into so-called top-sectors. This classification consists of nine main

categories plus one category "other" and 30 subcategories, this is further explained in section 2.1.

We explore the suitability of text mining by addressing two types of issues. Firstly, we investigate which settings and methods yield the best 'performance' in predicting the top-sector classification. Secondly, we investigate whether the performance of the predictions depend upon background variables of the businesses themselves.

From a societal point of view this problem is worth addressing, data mining applications such as text mining both can make the gathering of the data more efficient as well as enrich the available information (Hand, 1998; Cheung, 2012; Daas et al., 2015; Struijs et al., 2014; Hassani et al., 2014). Before this could be done it is important to find out if this automatic classification is possible with sufficient level of accuracy, so that the statistical quality can be guaranteed.

From a scientific point of view the problem is worth addressing since the benefits and problems in using big data getting a lot of attention from IT/organisational perspective but there is lack of attention from a statistical perspective (Daas et al., 2015). We know that text-mining techniques can be used for classification tasks and first experiments show promising results but development and evaluation of new methods are still needed while the success of the method is data specific (Hearst, 2003; Daas, 2012; Aphinyanaphongs et al., 2014; El-Halees, 2015; Tam and Clarke, 2015). Specifically, text mining has been applied successfully to automatic coding of industry and occupation based on text from open-answer survey questions, for instance recently by Gweon et al., (2017) and Thompson et al. (2012). It is interesting to find out to what extent such results can be replicated for texts extracted from websites.

The remainder of this paper is organised as follows. Section 2 presents the design of the study. Section 3 gives the results. Results are discussed in Section 4. Finally, section 5 concludes this report. In the appendix (section 6) some additional results are provided.

2. Approach

The dataset used in this research is created by combining, processing and splitting different datasets. Figure 1 illustrates how the final dataset was created. This will be explained in the rest of this section.

Section 2 is structured as follows. First, we give some background information on the case study (section 2.1). In section 2.2 we describe the 'Labels', the 'General Business Register (GBR) ' and the 'Population frame' that we used. Next, we explain how we have drawn a 'Sample' from this dataset, and how we transferred this dataset into an 'Anonymised sample'; both are described in section 2.3. In section 2.4 we describe how we obtained the 'Scraped dataset' and how the sampled data were pre-processed. In section 2.5 we describe which experiments we address in this study. Software is described in section 2.6 and parameter settings in section 2.7. Finally, in section 2.8 we explain how we evaluated the different experiments, including the split of the scraped dataset into a training-validation set' and a 'test set'.

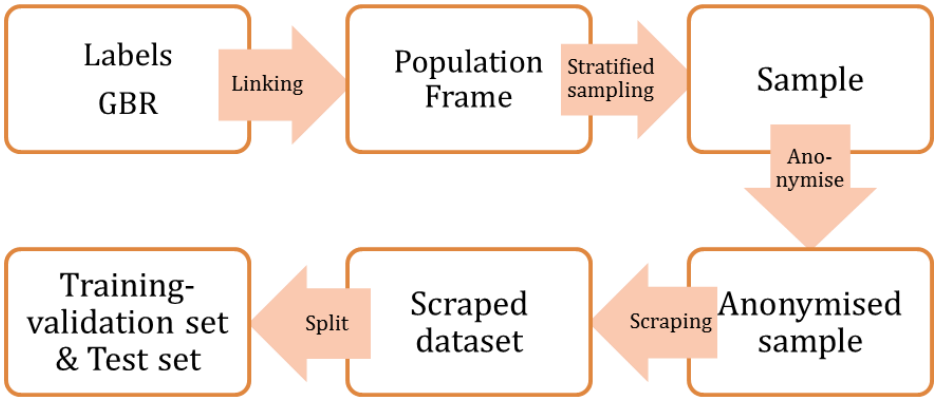

Figure 1 Process to create a training-validation and a test set

2.1 Case study

We used as a case study the annual monitor top-sectors (CBS, 2016). This monitor is based on a classification of nine economic 'top' sectors that are crucial for the Dutch economy. These top-sectors are 'Agriculture', 'Chemistry', 'Creative Industries', 'Energy', High tech systems and materials', 'Life sciences and health', 'Transportation and storage', 'Horticulture and raw materials' and 'Water'. Each top-sector in the annual monitor is divided in two to four sub-sectors, bringing the total to sub-sector 30 classes. In order to appoint all enterprises to a category, we introduced an additional category labelled by "other". A list of all top-sectors and their underlying sub-sectors is given in Table 1.

For the majority of the businesses, the top-sector classification follows from the main activity of businesses according to a certain grouping of their NACE code. In five of the top-sectors a small part of the businesses were appointed to a certain top-sector category on the basis of their membership of a certain trade organisation. These five top-sectors were 'Creative Industries', 'Energy', 'Transportation and storage', 'Horticulture and raw materials' and 'Water'. Note that the NACE code of an enterprise represents its main economic activity. However, a manually appointed top-sector class may concern a secondary economic activity.

2.2 Creation of the population frame

The population frame for the case study was created using two datasets:

– **General Business Register (GBR)**

The first dataset is a subset of the GBR containing 'businesses' that are known to be active in the Netherlands on 1-1-2017. Within the GBR there are different statistical unit types. In the present paper, we limit ourselves to the *enterprise* which we consider to be the statistical representation of a business. In the GBR a number of background variables are available: a unique identifier for each enterprise, the NACE code for economic activity, the URL (the web site address) and a classification of the size based on the number of employees. The GBR consist of 1.6 million enterprises.

– **Label sets for top-sectors and sub-sectors**

To identify the top- and sub-sector codes of the enterprises, a number of datasets from the year 2014 were available. First of all, there were nine datasets, for each top-sector one, containing a list of NACE-codes and their corresponding sub-sectors (29 in total) and top-sectors. Additionally there were five top-sectors for which a membership list was available edit with enterprises that belong to a certain top- and sub-sector on the basis of a trade organisation membership.

The top-sector classification is not completely disjoint. Some of the NACE codes belong to multiple top-sectors implying we have a multi-label classification problem.

The first step of the process was to link all those datasets together to create a population frame. In linking the datasets together four issues had to be taken into account:

- Firstly, since we want to classify enterprises on the basis of their website, only the enterprises from which the URL is known in the GBR have to be linked. From the 1.6 million enterprises in the Netherlands there are about 500.000 from which Statistics Netherlands (SN) knows the URL.
- Secondly, the top-sector classification includes a small and broad version of the category `Agriculture`. The broad version category overlaps with the small one but additionally it includes enterprises that are active in the production chain around food, such as enterprises engaged in the food transportation (CBS, 2016, p. 14–16). In the current research we used the broad definition.
- Thirdly, the label sets stem from the year 2014 and they need to be combined with a GBR population of 1 January 2017. The number of enterprises has increased from about 1.46 in 2014 to 1.6 million. To understand the

consequence of this time differences, we need to recall that for most of the population units the labels are based on NACE codes and for the other units on the trade organisation membership lists. This time-difference has no consequences for the first group, because the relation between label and NACE code did not change. Only some of the NACE codes had an extra digit in 2017 compared to 2014. This was solved by updating the lists with the relation between NACE code and top /sub-sector label.
For the second group, the units based on the trade membership list, the implications were slightly larger. We are actually interested in the trade membership list of 2017 (with their corresponding top-sector codes), but we only had the lists from 2014. This leads to a coverage error. First of all, the 2014 trade membership lists contained a total of about 2000 enterprises of which 249 could not be linked to the GBR of 1-1-2017. These concerned enterprises that existed in 2014 but ended somewhere between 2014 and 2017. Second, under coverage occurs because new enterprises have been started since 2014 , of which some probably became member of a trade organisation, but they are erroneously missing in our membership list.

- Fourthly, the enterprises that were not already appointed to a top-sector label, were appointed to a 10th top-sector category and a 30th sub-sector category 'other'.

The final population frame can be found in Table 1. Some enterprises belong to multiple top-sectors. Therefore, the sum of the number of enterprises over the separate top-sectors is larger than the total number of top-sector enterprises with a URL.

Table 1 Population frame

Category	Total	Membership list
Enterprises with URL	498 257	831
not in a top-sector (category "Other")	348 673	0
in a top-sector	149 584	831
Agriculture	22 435	0
Wholesale and retail Trade	14 418	0
Primary production	2 033	0
Manufacture of food products	3 753	0
Other	270	0
Chemistry	720	0
Manufacture of refined petroleum	6	0
Chemical industry	246	0
Manufacture of rubber and plastic	468	0
Creative Industries	84 315	59
Creative services	32 343	31
Cultural heritage	610	0
Art	30 827	0
Media and entertainment industry	20 535	28
Energy	641	482
Extraction of crude petroleum and gas	71	0
Sustainable energy	474	474
Related activities	96	8
High tech systems and materials	36 890	0
Manufacture of metal products	1 600	0
Manufacture of machinery	3 621	0
Manufacture of transport equipment	627	0
Other	31 048	0
Life sciences and health	795	0
Pharmaceutical	62	0
Manufacture of medical instruments	468	0
Research and development	265	0
Transportation and Storage	4 388	144
Transport	1 941	141
Warehousing and support activities	2 480	3
Horticulture and raw materials	2 411	54
Primary production	1 576	2
Other	835	52
Water	824	109
Construction of water projects	100	3
Building and repairing of ships and boats	638	28
Water collection, treatment and supply	56	48
Consultancy	30	30

2.3 The sample

We did not apply our text mining methods to all enterprises in the population, but we drew a sample. This was done due to practical reasons such as time and capacity. On average, it takes two minutes to scrape the homepage and the underlying layer of each webpage. This implies that it would take about two years to scrape websites of all Dutch enterprises when the robot server were to operate 24/7. In future, we may consider to use parallel processing to shorten the time needed to scrape the websites. That offers the opportunity to scrape a larger number websites. We could then use this larger set of websites to construct a so-called learning curve. A learning curve is a plot where the performance of a machine learning algorithm is plotted against the size of the training-validation set.

Our sample was drawn from the enterprises in the GBR that contain a URL. The sample size and the sampling design are described below.

2.3.1 Sample design

We aim to test the prediction of both top-sectors and sub-sectors, and we are interested to compare categories derived from the NACE code with those derived from membership lists.

When we would take a sample proportional to the population size of each category, we would obtain a very small number of units for some top-sectors and sub-sectors. Instead, we used a stratified sampling, where the current labels of the top- and sub-sectors were used as strata, as well as the property that an enterprise is on a membership list or not. We aim to achieve an overall good performance of text mining for all categories, so each class will have an equal weight (Weiss et al., 2010, pp. 214-215).

We are also interested to compare the performance for one-man-enterprises with larger enterprises. There was no need to stratify for that property, since both groups are well/represented in the population.

2.3.2 Sample size

Still the question remains what is the minimum sample size needed. This largely depends on the type of problem, the number of classes to predict and the classifier. A good approach for estimating the required sample size would have been to create a ‘representative’ learning curve (see above) for one of the categories of the classification. However, in the current study we limit ourselves to a first explorative analysis based on a limited overall sample size. Instead of using a learning curve we checked the literature on rules of thumb concerning the required sample size.

Stockwell and Peterson (2002) concluded that in general at least 20 examples per category are needed for a stable generalization performance. Furthermore Dumais

et. al. (1998) did a test with the SVM algorithm to explore the effect of the sample size on accuracy. A sample size of 70 led to 72.6% classification accuracy, a sample size of 350 to 86.2% accuracy, a sample size of 700 to 89.6% accuracy and a size of 7000 to 92% accuracy. From this study we conclude that a large sample size results in a (slightly) better performance but the effect decreases as more data are added.

Based on the above results, the following sampling set-up was selected. We first randomly select 70 enterprises (net sample size) within each sub-sector. That way we expected that we have sufficient training examples at sub-sector level. We then count the number of sampled units within each sub-sector that have a label based on a membership list. When that number is smaller than 20 (net sample size), we sampled additional units up to a minimum of 20, unless the population size of that group was smaller. In the latter case we sampled all population units in that group. The numbers of 70 and 20 were oversampled by 10%, because not all of the requested websites were actually retrieved. Main reasons for non-retrieval were: i) the website was no longer active, ii) the website was "for sale", and iii) we had an incorrect URL. We assumed that the non-retrievable websites are roughly evenly spread across the population. The final size of sample allocation is given in Table 2.

This sampling design means that at sub-sector level the sample is almost perfectly balanced. At top-sector level there is some imbalance because some top-sectors consist of multiple sub-sectors, but this imbalance stays within a reasonable range. This imbalance is greater when taking the multi-label instances character of the problem into account, since enterprises that are drawn from a certain sub-sector may also belong to another sub-sector. Therefore in the final sample dataset the enterprises were assigned to the labels in two ways:

- multi-label: all sub-sectors enterprises belong to according to the population frame;
- single-label: the sub-sector they were drawn from;

This makes it possible to experiment with how this characteristic influences the result.

After scraping we found that the actual non-response was 15%, so slightly larger than expected. We did not draw additional samples to correct for this. The net sample sizes for both single-label can be found in Table 3 and for multi-label in Table 4.

It is possible to correct for non-response and oversampling by adding a relative weight (w_h) to each unit in the text mining methods. This relative weight can be computed by dividing the response (r_h) with a constant (b) (let that number be 1) so that each sub-sector has exactly the same effective amount of training examples:

$$w_h = b/r_h = 1/r_h \quad (1)$$

Table 2 Sample allocation (gross sample)

Top-sector / sub-sector	Total	NACE	Member-ship list (M-list)	Over-sampling M- list
Agriculture	308			
Wholesale and retail Trade	77	77		
Primary production	77	77		
Manufacture of food products	77	77		
Other	77	77		
Chemistry	160			
Manufacture of refined petroleum	6	6		
Chemical industry	77	77		
Manufacture of rubber and plastic	77	77		
Creative Industries	352			
Creative services	99	77	0	22
Cultural heritage	77	77		
Art	77	77		
Media and entertainment industry	99	77	0	22
Energy	226			
Extraction of crude petroleum and gas	71	71		
Sustainable energy	77		77	
Related activities	78	70	6	2
High tech systems and materials	308			
Manufacture of metal products	77	77		
Manufacture of machinery	77	77		
Manufacture of transport equipment	77	77		
Other	77	77		
Life sciences and health	216			
Pharmaceutical	62	62		
Manufacture of medical instruments	77	77		
Research and development	77	77		
Transportation and Storage	174			
Transport	94	72	5	17
Warehousing and support activities	80	77	0	3
Horticulture and raw materials	174			
Primary production	79	77	0	2
Other	95	73	4	18
Water	244			
Construction of water projects	78	75	2	1
Building and repairing of ships and boats	96	74	3	19
Water collection, treatment and supply	56	8	48	
Consultancy	30		30	
Other	77	77		
Total	2239	1736	168	96

Table 3 Net Sample (single-label)

Top-sector / sub-sector	Total	Of which from membership list	Of which from one-man enterprise
Agriculture	266	0	94
Wholesale and retail Trade	66	0	21
Primary production	69	0	24
Manufacture of food products	67	0	19
Other	64	0	30
Chemistry	144	0	32
Manufacture of refined petroleum	6	0	1
Chemical industry	71	0	13
Manufacture of rubber and plastic	67	0	18
Creative Industries	305	41	152
Creative services	84	20	43
Cultural heritage	69	0	15
Art	66	0	42
Media and entertainment industry	86	21	52
Energy	186	74	37
Extraction of crude petroleum and gas	57	0	15
Sustainable energy	66	66	14
Related activities	63	8	8
High tech systems and materials	264	0	135
Manufacture of metal products	69	0	35
Manufacture of machinery	65	0	30
Manufacture of transport equipment	64	0	45
Other	66	0	25
Life sciences and health	179	0	57
Pharmaceutical	55	0	11
Manufacture of medical instruments	61	0	21
Research and development	63	0	25
Transportation and Storage	139	20	60
Transport	75	18	33
Warehousing and support activities	64	2	27
Horticulture and raw materials	151	24	41
Primary production	70	2	21
Other	81	22	20
Water	219	91	72
Construction of water projects	65	0	27
Building and repairing of ships and boats	28	28	3
Water collection, treatment and supply	76	19	33

Consultancy	50	44	9
Other	65	0	40
Total	1918	250	720

Table 4 Net sample (multi-label)

Top-sector / Sub-sector	Total	Of which from membership list	Of which from one-man enterprise
Agriculture	310	0	108
Wholesale and retail Trade	87	0	25
Primary production	80	0	40
Manufacture of food products	69	0	24
Other	74	0	19
Chemistry	180	0	44
Manufacture of refined petroleum	6	0	1
Chemical industry	74	0	13
Manufacture of rubber and plastic	100	0	30
Creative Industries	305	41	152
Creative services	83	19	43
Cultural heritage	69	0	15
Art	66	0	42
Media and entertainment industry	87	22	52
Energy	212	86	38
Extraction of crude petroleum and gas	57	0	15
Sustainable energy	77	77	14
Related activities	64	9	8
High tech systems and materials	399	0	180
Manufacture of metal products	69	0	35
Manufacture of machinery	163	0	51
Manufacture of transport equipment	103	0	40
Other	64	0	45
Life sciences and health	192	0	62
Pharmaceutical	55	0	11
Manufacture of medical instruments	74	0	26
Research and development	63	0	25
Transportation and Storage	140	21	60
Transport	76	19	33
Warehousing and support activities	64	2	27
Horticulture and raw materials	159	24	42
Primary production	73	2	21
Other	85	22	21
Water	222	94	72
Construction of water projects	65	0	27

Building and repairing of ships and boats	31	31	3
Water collection, treatment and supply	75	19	33
Consultancy	51	44	9
Other	65	0	40

However because the level of non-response was roughly equal between sub-sectors this was not necessary.

Finally, we transformed the sample into a confidential sample. Due to judicial restrictions, the security of the data has to be ensured. So, after the sample was drawn identifying characteristics such as the unique identifier was removed and replaced by a local identifier for the sample. So was ensured that no sensitive information was lost during web scraping. See Table 2.

2.4 Scraping and pre-processing the website data

We scraped the webpages of the sampled enterprises using a robot server at SN. The resulting data were stored using the local identifier into a database at SN.

We first needed to decide which parts of the website were needed to obtain useful information for text mining. We judged this "usefulness" by checking to what extent the retrieved words coincided with the words in our dictionary. We started with a preliminary manual assessment in the top-sector 'Agriculture', where the header, the homepage plus one additional layer was scraped. The assessment showed that either the website returned words that were also found in our dictionary or the website did not respond or there was no useful information on the website (for example a message that the website was for sale). The analysis showed that scraping only the headers returned an insufficient amount of text (see Figure 2) for the feature selection. Therefore, for the remainder of our paper, we scraped the header, the homepage plus one additional layer. In future research, we may further investigate which parts of the website are most useful for predicting economic activity (see discussion).

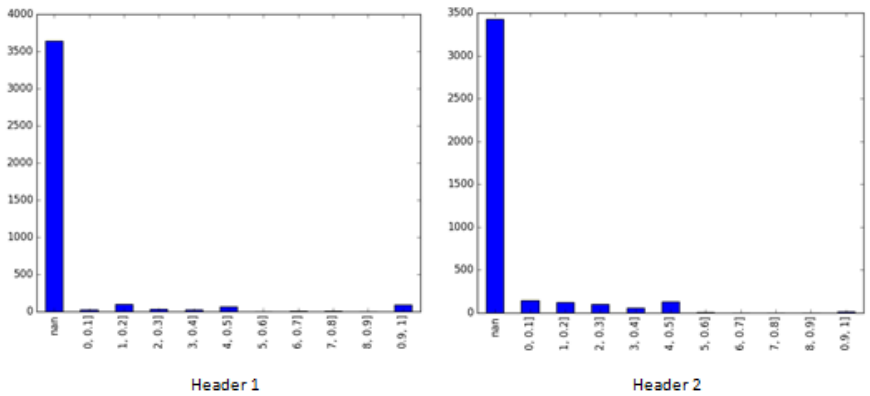

Figure 2 Frequency distribution of the number of words in the dictionary filter for the fraction of the webpage headers that are scraped, for header (level) 1 and 2

At the same time, the assessment revealed some difficulties. The language of the websites was very divers. By analysing the website language in Python, using the package Langdetect[1], a total of 30 languages was found. The Dutch language was used in about 70% of the cases. English was the second most common language. Furthermore there was a small fraction of French, German and South-African websites. That we found so many different languages for top-sector enterprises is not surprising since the Department of Economic Affairs defines 'international export orientation' as one of the characteristics of an enterprise belonging to a top-sector (CBS, 2014).

In our study, we decided to limit ourselves to Dutch websites. This way, we only needed Dutch words as features. We used Langdetect to select websites with Dutch as a language. For each website, a probability distribution of the detected languages was obtained. We selected a website as being "Dutch" when the sum of the probabilities of the languages *Dutch* and *South-African* was at least 90%. We assumed that the language of websites classified as South-African were in fact Dutch websites.

To prepare the data set for analysis a number of general pre-processing steps were taken. First step was the cleaning of the text. The HTML related content was removed along with punctuation, whitespaces and numbers. Furthermore, all the text was transformed from uppercase to lowercase. Additionally, to clean the text stop words were removed with a list of 240 Dutch stop words.

Next, the words were stemmed with a Dutch stemmer. Stemming is a way to transform derivations of a word into the same form. This is done by a rule-based algorithm that transforms the different terms to the root or another stem if the term has different grammatical forms (Porter, 2001). Finally, the cleaned and stemmed website content was transformed in a format that can be interpreted by the algorithm. This was done using the bag of words assumption (see 2.3.2.1.): transforming the website into feature vectors with the counts of words for each website. In the current study, we restricted the words entering the feature vectors to single words (unigrams). The use of n-grams is left for future research (see section 4).

As explained in the introduction, we explore the suitability of text mining by addressing two types of issues:

- which settings and methods yield the best 'performance' in predicting the top-sector classification;
- To what extent does the performance of the predictions depend upon the complexity of the websites and of the enterprises.

The following settings and methods were tested:

[1] Langdetect 1.0.7: https://pypi.python.org/pypi/langdetect

– **Word weighing (TF-IDF)**

Two word weighting methods were compared: term frequency (TF: the normal word count) and word count as expressed relative to its frequency in document inverse document frequency (inverse document frequency, TF-IDF), see Zhang et al. (2011, p. 2760). Since the IDF weighting is used to find the most relevant words it is the question whether this is useful in combination with a dictionary filter (See "Feature selections").

– **Multi-label versus Single-label**

The problem at hand is a multi-label problem: each enterprise may have multiple labels. Because only one-third of the enterprises have two or more labels, we could also simplify the situation to a single-label problem. Moreover, each enterprise is coded in the GBR with a single main activity (NACE code). Results from the single-label approach may give insight into the potential of text mining to predict NACE codes.

In case of multi-label prediction, each classifier predicts one label at a time, through a one versus rest approach (OVR).

– **Dictionary filters**

Another very important variation this research tests is the effectiveness of three sets of different feature selections methods: a dictionary filter, automatic feature selection and the pooled set of both. The dictionary filter was based upon a set of terms that is being used at SN for each domain to manually classifying economic activity of enterprises. To make sure the words in the dictionary match the scraped words on a website also the used lexicon was stemmed as well. We also refer to this dictionary as the NACE filter. This knowledge-based dictionary filter was compared with an automatic feature selection dictionary (K-best dictionary) of the same size (about 4200 features) that is automatically selected by selecting the features with the most variance. As a third variation the words found in both approaches (NACE dictionary and automatic feature selection) were also pooled.

This knowledge-based dictionary filter for top-sector is related the manual coding of NACE codes and therefore will be referred to as NACE-dictionary. The assumption behind test is that the words that human use to classify an enterprise for top-sectors will also be the words an algorithm helps to distinguish classes. In sentiment analysis the use of a dictionary filter is already common practice as a feature selection method: selecting only words related to emotion boosts classification performance (Kouloumpis et al., 2011).

– **Classifiers**

This concerned the following five classifiers: *k*-Nearest Neighbour (KNN), Random Forrest (RF), Naïve Bayes (NB), Support Vector Machines (SVM) and Logistic Regression (LR). With this set of classifiers a variety of different mathematical ways to make decision is explored to measure the effect of the different configurations. Apart from the single classifiers, we also test an ensemble method, namely a voting classifier. The latter is further explained in section 3.4.

The following elements of "complexity" of the websites at hand were tested:

- **Scraped parts of the web site**

We compared variations in the parts of the website content that was scraped to get an idea which part of the website contains the most relevant information about economic activity. We varied scraping:

- the homepage
- the homepage plus one deeper layer of webpages.

As there are examples of private companies who predict enterprise sector solely on the homepage the process would be a lot more efficient when only one page per website has to be scraped (Rigter, 2017). Another variation that could be made is between the body text and the different headers on the website. However as already illustrated (see Figure 1) this variation probably would not yield much success as the input content was too limited. Therefore this variation was eventually not included in the research.

- **Enterprise size**

We compared the text mining performance of one-man enterprises versus larger enterprises.

- **Label allocation**

We compared the text mining performance of enterprises whose top-sector class is based on their NACE code versus enterprises whose top-sector class is based upon the membership list.

2.5 Design of experiments

We tested a large number of combinations between the different variations given in section 2.4. These results can be found in Roelands (2017). Here, we limit ourselves to presenting a number of experiments to evaluate the effect of the different variations. The result is summarised in Table 5.

The rationale behind this experiments is the following. We first defined which of the variations we considered to be the default setting. Next we varied one of the components at a time with respect to this default setting. The default setting was:

- use the setting with "optimal performance" (see section 2.8) for the word weighting method (TF versus TF-IDF). This optimum can be found by a grid-search approach were also the other parameters of the text mining are varied;
- use a multi-label prediction, since the problem is multi-label by nature;
- use the "optimal" dictionary filter. The optimum is found by manually comparing the results of the three different dictionary variations. The optima concerns the following combinations of classifiers and dictionaries: KNN - K-best, RF - Intersect, NB - Pooled, - SVM - Intersect, LR - Intersect;
- give the results of all five classifiers;

- predict both top-sector and a sub-sector level;
- use the results of all enterprises (thus one-man enterprises and larger enterprises, enterprises with label based on NACE code and those based on the membership list)
- use the scraping results of home page and one underlying layer.

For all the experiments we give results at both top-sector as sub-sector level as the degree of detail in the classes might give other results.

Finally the variations in the characteristics that possibly influence complexity are evaluated on for the classifier that that gives the best outcome for the default setting. The complexity evaluation was only conducted at top-sector level because otherwise the number of training samples was too low.

Table 5 Experiment configuration

	Variations					
Experiments		**Word Weighting**	**Label**	**Dictionary filter**	**Classifier**	**Detail prediction**
	Word weighting	TF/TF-IDF	Multi	Optimal[2]	All	Top-sector & Sub-sector
	Label	Grid search	Single/ Multi	Optimal	All	Top-sector & Sub-sector
	Dictionary filter	Grid search	Multi	NACE / K-best / Pooled / Intersect	All	Top-sector & Sub-sector
	Classifiers	Results analysed over the experiments				Top-sector & Sub-sector
	Complexity Evaluation	Best	Best	Best	Best	Top-sector

2.6 Software

We used the following software:

- **Elastic Search:**

Elastic Search is an open source search engine used to search the webpages stored as documents retrieved from the web with a robot server. Elastic Search enabled us to search and navigate through the HTML interface stored webpages to clean the text and extract the useful terms. This was done by two dictionaries reducing the 240.000 features long vector to the 4200 selected features for each dictionary.

- **Python**:

2 The optimum is not found through a grid search but manually picked on the basis of the results of the dictionary filter experiment. The optima concerns the following combinations of classifiers and dictionaries: KNN - K-best, RF - Intersect, NB - Pooled, - SVM - Intersect, LR - Intersect

Most of the work was implemented with Python 3.5 software. For the machine learning package Scikit-learn version 0.17 (Pedregosa *et al.*, 2011.) was used. A pipeline setup with the TF-IDF transformer and a grid search was conducted to find the best parameters for each classifier (see Table 6). When predicting a multi-labelled problem the classifiers were implemented via one-vs-rest (OVR) classifier, using the different classifiers as an estimator. The majority voting learning ensembles that were shortly touched upon were implemented via a Voting Classifier.

2.7 Parameter settings

The parameters for the grid search of the classifiers were set fairly broad, but not too wide (see Table 6). An explanation of the parameter can be found in Scikit-learn (Pedregosa et al., 2011). The settings were based on some first explorations, to ensure that the optimum was found within the set of grid search parameters. Nonetheless, our results should be understood as the best classifier given these grid-search parameters settings, since we could impossibly test all possible parameters. For the voting classifier the parameters are chosen based on the outcomes of earlier experiments, since it would take much computation time to include all the parameters in the grid search.

Table 6 Grid search parameter grid for Python implementation

Classifier	Parameters
KNN	Number of neighbours in KNN (n_neighbours): [1,3,5,7,9,11,13,15,17,19] metric: [‘cosine’, ‘euclidean’, ‘minkowski’]
RF	min_samples_split: [2,5, 10, 20, 30,40,50,60,70,80,90,100] bootstrap: [True, False] criterion: ["gini", "entropy"] number of trees in the forest (n_estimators): [5, 10, 15, 20, 25, 30]
NB	alpha: [1, 0.1, 0.01, 0.001, 0.0001, 0.00001, 0] fit_prior: [True, False]
SVM	C: [0.01, 0.1, 1, 2, 4, 8, 10, 20, 50] gamma: [0.0001, 0.001, 0.01, 0.1, 1, 2, 10] kernel': ['linear', 'poly', 'rbf', 'sigmoid']
LR	penalty : ['l2', 'elastic net'] number of iterations used to find the parameter estimates (n_iter): [5, 10, 50, 100] alpha: [0.001, 0.0001, 0.00001, 0.000001]

2.8 Evaluation

The different experiments were evaluated in the following way.

2.8.1 Construction of a training and test set

To evaluate performance, the sample was split into a separate training-validation set and a test set. We used a ratio of 80/20, meaning there were roughly 300 enterprises

(10 per sub-sector) in the test set to evaluate the performance on. The training-validation set, containing roughly 30 enterprises per sub-sector, was used to train and tune the parameters of the text mining method. A five-fold cross validation was used to split the training-validation set into a training and validation part to prevent overfitting.

2.8.2 Evaluation metrics

The evaluation of classification predictions is presented in the form of a confusion matrix (see Table 7) . In a confusion matrix one counts the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). In addition the confusion matrix gives the margin of the total number units that in reality are positive (PO) and those that are negative (NE). Based on these counts, the following evaluation metrics were considered:

- accuracy (A):

$$A = \frac{TP + TN}{PO + NE} \tag{2}$$

- precision (P), also referred to as positive prediction value and specificity:

$$P = \frac{TP}{TP + FP} \tag{3}$$

- recall (R), also referred to as hit rate, true positive rate and sensitivity:

$$R = \frac{TP}{TP + FN} \tag{4}$$

- F1 scores, the harmonic mean of precision and recall:

$$F1 = \frac{2 * PR * R}{PR + R} \tag{5}$$

Table 7 Confusion matrix

	Predicted		
True		Positive	Negative
	Positive (PO)	True Positive (TP)	False Negative (FN)
	Negative (NE)	False Positive (FP)	True Negative (TN)

The evaluation metrics were used for both the validation and the test set. For the test set, we used all four measures to give a broad overview of the performance of the methods. For the validation set, a single evaluation metric was selected that was used to optimise the tuning parameters of the text mining methods. The most basic evaluation metric is accuracy. The downside of using accuracy for parameter tuning is that it pushes the algorithm to behave like a trivial rejecter, assigning documents where the confidence in the decision is low to the majority class (Sebastiani, 2002, p. 34). The sample design should help to overcome this behaviour, but there is still some imbalance in the dataset. We used the F1 score instead of the accuracy, which yields a balance between precision (the percentage where a certain label is predicted correct) and recall (the percentage where a certain label should be predicted is

correct). Therefore, precision is referred to as specificity and recall as sensitivity (Powers, 2011, p.38).

Note that for the evaluation of the performance of the multi-label classification problem, some adjustments were needed. We first consider each label separately, and count the number of TP, FP, TN and FN. Next, we combined the outcomes for all of the labels. We decided to use the micro-average method, With micro-averaging the enterprises that are in the TP, FP, TN and FN part of the confusion matrix are added up at micro level. The latter was being done because the sample was fairly balanced. In the case of multi-label classification each correct prediction for a class is seen as TP, each incorrect decision for a class is seen as FP and each missing prediction as FN (Van Asch, 2013). Furthermore the confusion matrix for the full set of labels gives insight in which mistakes are made. The latter can then be used to train and to tune the algorithm further.

We need a criterion in order to decide whether the text mining is a potentially useful method in case of industry codes form enterprise websites. One option is to compare the performance with a base-line method, see for instance Thompson et al. (2012). Another option in our situation is to compare the performance of the classifiers with a majority vote method: always selecting the most frequently occurring category. However, this is not very demanding criterion. The results need to be of sufficient quality before automatic classification can be used to replace manual classification.

A natural alternative is to use a minimal accuracy. Williams (2006) tested an Automatic Coding by Text Recognition (ACTR) of statistics Canada applied to automatic coding of economic activity. He aimed to achieve a quality threshold of "7.5 on a scale of 0 to 10" which is too vague for us to be useful. Thompson et al. (2012) investigates automatic coding of industry and occupation from answers obtained in survey sampling. They aimed to achieve an accuracy of a new system with a maximum error rate equal to that in manual coding: 5%. In many cases one aims to develop a system where easy-classify cases are coded automatically while hard to-classify-cases are classified manually; the distinction between the two is based on a certain threshold. In this context, the production rate is the fraction of all cases that is classified automatically. Chen et al. (1993) provides an example of the trade-off between accuracy and production rate. Thompson et al. (2012) achieved a production rate of 55% at a 5% error rate. Jung (2008), studying automatic coding of industry and occupation in Korea, achieved a production rate of 83% at 98% accuracy.

In the present study we explore the potential of automatic text mining, where we limit ourselves to fully automatic classification, as a 100 % production rate. An example of the performance at 100% production rate for a number of new text mining methods for occupation coding is given in Gweon et al. (2017). Their study concerned 390 different occupational codes. At fully automatic classification, their weakest performing machine learning method yielded an accuracy of just below 53% while their best performing method yielded a 65% accuracy (Figure 3 in Gweon et al. (2017). We will compare our results with this "benchmark". In our situation, the

measure precision should be large enough, since automatic classification by text mining mainly serves as a complementary method to manual classification. So, when a category is automatically assigned, there should be a high probability that the assignment is correct.

3. Results

The results presented in the current section are a selection of the full set of experiments that have been computed. The remaining results are given in section 6.

The tables in this section and in the Appendix should be understood in the following way. The pre-processing variation applied can be found in the variation column. The optimal parameters found by the grid search are displayed in the parameter column. Further the cross-validation score (F1-score) is given (for the validation set) as well as the micro averaged accuracy (A) precision (PR), recall (R) and F1-score (F1) of the separate test set. This score should be compared with the following majority baseline:

Label	Class	Accuracy, precision, recall, F1-score
Top-sector, Multi label	Agriculture	233/1432 = 0.163
Top-sector, Single label	Agriculture	202/1258 = 0.160
Sub-sector, Multi label	Agriculture – primary production	72/1432 = 0.050
Sub-sector, Single label	Transportation and storage – transportation	69/1258 = 0.055

3.1 The effect of different word weighting methods

As described in the theoretical framework when a bag of words assumption is applied a document can be represented by the TF or the TF-IDF weighting method. The result of this comparison for top-sectors can be found in Table 8 and the results for sub-sector can be found in Table 19. These tables show the effect of TF-IDF weighting for each classifier.

What stands out when comparing performance for both top-sectors as well as for sub-sectors is that the differences between TF and TF-IDF for three of the five classifiers (NB, SVM, LR) are very small and also can vary between the test set and validation set. In these cases, the use of TF-IDF weighting results in a higher precision but a lower recall, resulting in a F1-score that is more or less the same. For the KNN classifier the difference is clearer: the use of TF-IDF leads to an increase in F1-score of 13 (top-sectors) and 7 (sub-sectors) percentage points. The Random Forrest at the first sight seems to prefer the TF over the TF-IDF. However, this is solely the case for top-sector when studying the result for sub-sector in Table 19 it appeared that with a small margin TF-IDF is preferred. For the remainder of this paper the word weighting was included in the grid search as one of the variables to tune. The grid search results (see the 'use_idf' parameter) confirm the conclusion that as for most classifiers difference between TF and TF-IDF is small, given the specific experiment the choice for TF or TF-IDF weighting differed.

Table 8 Evaluating the effect of TF-IDF weighting for predicting top-sectors

Classifier	Variation	Parameters	Validation score (F1)	A	P	R	F1
KNN	TF	n_neighbours = 1 metric = Cosine	0.468	0.328	0.43	0.43	0.42
	TF-IDF	n_neighbours = 3 metric = Cosine	0.566	0.415	0.67	0.49	0.55
RF	TF	Criterion = 'gini' min_samples_split = 5 n_estimators = 50	0.569	0.444	0.82	0.47	0.58
	TF-IDF	Criterion = 'gini' min_samples_split = 5 n_estimators = 50	0.553	0.444	0.82	0.47	0.57
NB	TF	Alpha = 0.00001 fit_prior = False	0.640	0.477	0.62	0.67	0.64
	TF-IDF	Alpha = 0.001 fit_prior = True	0.656	0.448	0.78	0.53	0.63
SVM	TF	C = 8 Gamma = 0.1 Kernel = 'linear''	0.541	0.389	0.68	0.49	0.57
	TF-IDF	C = 8 Gamma = 0.001 Kernel = 'linear'	0.571	0.437	0.70	0.54	0.60
LR	TF	Alpha = 0.0001 n_iter = 50 penalty = 'l2'	0.562	0.448	0.67	0.53	0.58
	TF-IDF	Alpha = 0.00001 n_iter = 5 penalty = 'elasticnet'	0.569	0.472	0.67	0.56	0.60

3.2 The effect of one or more labels

We compare the effect of multi-label with single label prediction. Although the problem is in fact multi-label and this is also the standard set-up, for this specific experiment the single label prediction is applied (see section 2.3 for an explanation) and compared with multi-label prediction. The results of this experiment comparing the effect for each classifier can be found in Table 9 for predicting top-sectors as the results for predicting sub-sectors can be found in Table 10.

Table 9 Evaluating the effect of multi label and single label for top-sectors

Classifier	Variation	Parameters	Validation score (F1)	A	P	R	F1

KNN	Single label	Use_idf = True n_neighbours = 13 metric = Cosine	0.524	0.564	0.63	0.56	0.56
	Multi label	Use_idf = True n_neighbours = 3 metric = Cosine	0.566	0.415	0.67	0.49	0.55
RF	Single label	Use_idf = False Criterion = 'gini' min_samples_split = 10 n_estimators = 50	0.629	0.619	0.62	0.62	0.59
	Multi label	Use_idf = False Criterion = 'gini' min_samples_split = 5 n_estimators = 50	0.554	0.421	0.84	0.46	0.57
NB	Single label	Use_idf = True Alpha = 0.001 fit_prior = True	0.636	0.647	0.65	0.65	0.65
	Multi label	Use_idf = True Alpha = 0.001 fit_prior = True	0.656	0.448	0.78	0.53	0.63
SVM	Single label	Use_idf = True C = 1 Gamma = 0.01 Kernel = 'linear’	0.500	0.536	0.59	0.54	0.52
	Multi label	Use_idf = True C = 8 Gamma = 0.001 Kernel = 'linear'	0.571	0.437	0.70	0.54	0.60
LR	Single label	Use_idf = True Alpha = 0.0001 n_iter = 50 penalty = 'l2'	0.547	0.547	0.56	0.55	0.54
	Multi label	Use_idf = True Alpha = 0.0001 n_iter = 5 penalty = 'elasticnet'	0.575	0.425	0.74	0.50	0.59

To start with, it is good to point out that at top-sector level the F1 scores of single- and multi-label classification do not differ much. However, when further analysing the results there are three differences that stand out. First, the accuracy of the classifiers is smaller with multi-label compared to with single-label. This indicates that in case of multi-label it is harder to predict all the labels from an instance completely correct. The classifier might predict one of the multiple labels of an instance correct but not all labels. Second, multi-label prediction leads to a larger difference between precision and recall than single-label. As the multi-label prediction is run in a one versus the rest configuration, the algorithm makes a decision for each class by comparing it with all other classes. The multi-label prediction is more carefully

assigning a label but is also missing labels more often. The single-label prediction on the other hand is forced to make a decision, resulting in a lower precision but a higher recall than multi-label prediction.

Table 10 Evaluating the effect of multi label and single label for sub-sectors

Classifier	Variation	Parameters	Validation score (F1)	A	P	R	F1
KNN	Single label	Use_idf = True n_neighbours = 13 metric = Cosine	0.382	0.415	0.45	0.41	0.40
	Multi label	Use_idf = True n_neighbours = 3 metric = Cosine	0.414	0.290	0.60	0.32	0.39
RF	Single label	Use_idf = True Criterion = 'gini' min_samples_split = 30 n_estimators = 50	0.443	0.437	0.47	0.44	0.42
	Multi label	Use_idf = True Criterion = 'gini' min_samples_split = 1 n_estimators = 5	0.259	0.179	0.56	0.21	0.33
NB	Single label	Use_idf = False Alpha = 0.001 fit_prior = True	0.472	0.477	0.48	0.48	0.48
	Multi label	Use_idf = True Alpha = 0.001 fit_prior = True	0.434	0.253	0.55	0.36	0.42
SVM	Single label	Use_idf = True C = 10 Gamma = 0.1 Kernel = 'linear''	0.342	0.361	0.42	0.36	0.35
	Multi label	Use_idf = True C = 8 Gamma = 0.001 Kernel = 'linear'	0.395	0.277	0.56	0.36	0.42
LR	Single label	Use_idf = True Alpha = 0.0001 n_iter = 5 penalty = 'elasticnet'	0.371	0.385	0.40	0.38	0.37
	Multi label	Use_idf = True Alpha = 0.00001 n_iter = 100 penalty = 'l2'	0.401	0.277	0.52	0.37	0.42

At sub-sector level the single-label prediction always resulted in a higher F1 score than the multi-label prediction for a given classifier. At top-sector level it varied whether the single-label or the multi-label configuration had a higher F1 score.

3.3 The effect of different methods to select a dictionary

We compared four variations of feature selection: NACE dictionary, K-best dictionary (see section 2.4), Pooled dictionary and Intersection dictionary. To explore whether the knowledge-based NACE dictionary was suitable for the task, we analysed the uniqueness of the selected terms for each top-sector. We computed the jacquard index which is the intersection relative to the union of the word sets between two classes, at top-sector and sub-sector level (see Table 20). For most of the classes, the terms in the dictionaries did not have a large overlap with those of other classes. The index showed that there are 22 combinations that have 10% or more overlap and just 5 combinations with 15% or more overlap. Moreover, there is a relatively small overlap between the sub-sectors that belong to the same top-sector, indicating that predicting at a more detailed level may not be more difficult. However as previous results already indicated comparing the performance on top-sector level and sub-sector level confirmed that predicting sub-sector with this dictionary is still a difficult task.

To investigate to what extent new information is added by pooling the dictionaries the jacquard index was used to measure the overlap. There are 655 words in the intersection which is 7.8% of the combined dictionary length. This motivated us to construct a fourth dictionary containing the intersection of both sets: that set contains words that are useful from a knowledge-based perspective and from an automatic feature selection perspective.

The difference in performance between the knowledge-based NACE dictionary and the K-best dictionary depends on the classifier, whether the prediction is at top-sector level (Table 11) or at sub-sector level (Table 21). At top-sector level for instance, the K-best dictionary gives better results for the KNN and NB classifier while for RF, SVM and LR the NACE dictionary gives better results. The confusion matrices reveal some (marginal) differences in the type of mistakes the classifiers make (not shown).

The pooled dictionary did not prove to be very successful. For the NB and RF classifier the pooling of the dictionary gives a slightly better F1-score where for the other classifiers the pooled dictionary F1-score was in between the F1-score of the NACE and K-best dictionary. The intersection dictionary appeared to have more effect. This dictionary yields the best F1-score for the RF, SVM and LR classifier, both at top-sector (Table 11) and at sub-sector level (Table 21).

Table 11 Evaluating the effect of the dictionaries for predicting top-sectors

Classifier	Variation	Parameters	Validation score (F1)	A	P	R	F1
KNN	NACE	Use_idf = True n_neighbours = 3 metric = Cosine	0.450	0.257	0.51	0.31	0.38
	K-best	Use_idf = True n_neighbours = 3 metric = Cosine	0.566	0.415	0.67	0.49	0.55
	Pool	Use_idf = True n_neighbours = 5 metric = Cosine	0.490	0.295	0.66	0.35	0.44
	Intersect	Use_idf = True n_neighbours = 3 metric = Cosine	0.453	0.341	0.62	0.40	0.48
RF	NACE	Use_idf = False Criterion = 'gini' min_samples_split = 1 n_estimators = 5	0.442	0.286	0.71	0.35	0.46
	K-best	Use_idf = True Criterion = 'gini' min_samples_split = 1 n_estimators = 5	0.416	0.245	0.71	0.33	0.44
	Pool	Use_idf = False Criterion = 'gini' min_samples_split = 5 n_estimators = 5	0.467	0.270	0.72	0.37	0.48
	Intersect	Use_idf = False Criterion = 'gini' min_samples_split = 5 n_estimators = 50	0.554	0.421	0.84	0.46	0.57
NB	NACE	Use_idf = False Alpha = 0.0001 fit_prior = False	0.632	0.452	0.56	0.65	0.60
	K-best	Use_idf = True Alpha = 0.001 fit_prior = True	0.611	0.432	0.75	0.52	0.61
	Pool	Use_idf = True Alpha = 0.001 fit_prior = True	0.656	0.448	0.78	0.53	0.63
	Intersect	Use_idf = True Alpha = 0.00001 fit_prior = False	0.584	0.393	0.55	0.74	0.62

Table 11 (Cont.)

Classifier	Variation	Parameters	Validation score (F1)	A	P	R	F1
SVM	NACE	Use_idf = True C = 8 Gamma = 0.01 Kernel = 'linear''	0.563	0.390	0.67	0.49	0.56
	K-best	Use_idf = False C = 8 Gamma = 0.001 Kernel = 'linear'	0.557	0.340	0.63	0.45	0.52
	Pool	Use_idf = False C = 8 Gamma = 0.1 Kernel = 'linear'	0.571	0.352	0.68	0.46	0.54
	Intersect	Use_idf = True C = 8 Gamma = 0.001 Kernel = 'linear'	0.571	0.437	0.70	0.54	0.60
LR	NACE	Use_idf = True Alpha = 0.0001 n_iter = 100 penalty = 'l2'	0.567	0.376	0.67	0.45	0.54
	K-best	Use_idf = True Alpha = 0.0001 n_iter = 5 penalty = 'elasticnet'	0.562	0.344	0.71	0.42	0.52
	Pool	Use_idf = True Alpha = 0.00001 n_iter = 5 penalty = 'elasticnet'	0.586	0.336	0.74	0.41	0.52
	Intersect	Use_idf = True Alpha = 0.0001 n_iter = 5 penalty = 'elasticnet'	0.575	0.425	0.74	0.50	0.59

3.4 The effect of different Classifiers

At top-sector level, the best score of the least performing classifier (KNN) is 0.55 while the best score of the best performing classifier (NB) is 0.63. The NB classifier proofed to be a robust performer over the whole range of experiments, often being the best classifier. The RF, SVM and LR classifiers are more sensitive to the variations tested, for instance for the set of features that are used. There was a clear difference in performance of the classifiers between top-sector and sub-sector level. The best

method for top-sectors has an F1-score of 0.63 while the best method for sub-sectors has an F1-score of 0.44.

To improve the performance, the effect of learning ensembles was studied. We limited this to the intersection dictionary since that dictionary on average gave the highest F1-scores. As a learning ensembles we used a Voting Classifier with a simple even weighted voting mechanism. The five different classifiers were included in a grid search. The results of the best preforming combinations of classifiers are given in Table 12.

Table 12 Voting classifier implementation for predicting top-sectors

Variation	Parameters	Validation score (F1)	A	P	R	F1
Intersection dictionary	Use_idf = True estimators = NB, LR, RF	0.633	0.512	0.80	0.58	0.66

The results show that implementation of the voting classifier lead to a small increase, about 3 percentage points, of the F1-score. For sub-sectors the implementation of the voting classifier did not result in a higher F1-score. Table 22 and Table 23 in the appendix give an overview of the classification performance of the best classifiers for each class. For top-sectors, the score for precision is stable except for the class 'other' while the score for recall show somewhat larger fluctuations. For sub-sectors, the values of the different measures fluctuate considerably. Larger samples are needed, before we can draw clear conclusions whether certain classes are more difficult to predict than others.

3.5 Evaluating complicating characteristics

The influence of the complicating characteristics on the classifier results are tested at top-sector level on the best results based on the previous three experiments (word weighting, multi- / single label and the dictionary filter). We concluded (see Table 11) that the NB classifier gave the best results, with TF-IDF weighting and the pooled dictionary. We now investigate the effect of the size of the enterprise, of enterprises labelled on the basis of their NACE code or based on a membership list and website complexity.

3.5.1 The size of the enterprise

We compared the classification performance on one-man enterprises (Table 13) with those enterprises that have more than one employee (Table 14). The F1-score was 8 per cent points larger for websites of enterprises with more than one employee than for websites of one-man enterprises. Since the precision of both groups is roughly equal the difference is due to the difference in the recall score. A manual assessment of a sample of websites revealed that websites of one-man-enterprises are often more compact than those larger enterprises, resulting in a smaller set of words. Still, the results should be interpreted with some caution. The final column (labelled:

"support") in Table 13 and Table 14 contains the number of enterprises in each class. In the top-sectors 'energy' and 'horticulture and raw materials' no cases of one-man enterprises were present in the test set. It needs to be seen if these findings still hold with larger number of enterprises.

Table 13 classification report one-man-enterprise
accuracy on the test set: 0.42

Top-sector	Precision	Recall	F1-score	Support
Other	0.00	0.00	0.00	6
Agriculture	0.90	0.60	0.72	15
Chemistry	0.50	0.20	0.29	10
Creative Industries	0.81	0.59	0.68	22
Energy	0.00	0.00	0.00	0
High tech systems and materials	0.92	0.55	0.69	22
Life sciences and health	1.00	0.50	0.67	2
Transportation and storage	1.00	0.44	0.62	9
Horticulture and raw materials	0.00	0.00	0.00	0
Water	0.75	0.43	0.55	7
Average / Total	0.77	0.46	0.57	95

Table 14 classification report enterprise with more than 1 employee
accuracy on the test set: 0.46

Top-sector	Precision	Recall	F1-score	Support
Other	0.20	0.25	0.22	4
Agriculture	0.74	0.53	0.62	32
Chemistry	0.71	0.50	0.59	10
Creative Industries	0.90	0.76	0.83	25
Energy	0.69	0.64	0.67	14
High tech systems and materials	0.77	0.55	0.64	42
Life sciences and health	1.00	0.71	0.83	17
Transportation and storage	0.67	0.31	0.42	13
Horticulture and raw materials	0.70	0.44	0.54	16
Water	0.85	0.61	0.71	18
Average / Total	0.78	0.57	0.65	191

3.5.2 The source of the labels

Enterprises that are labelled on the basis of a membership list had a lower F1-score as averaged over the different top-sectors, that those labelled on the basis of their main economic activity (NACE code) (see Table 15 and Table 16). Results suggest that membership list enterprises are harder to predict. This is a preliminary result, because the number of enterprises with a membership list in the test was small.

Table 15 classification enterprises based on membership list

Accuracy on the test set: 0.36

Top-sector	Precision	Recall	F1-score	Support
Other	0.00	0.00	0.00	0
Agriculture	1.00	1.00	1.00	1
Chemistry	0.00	0.00	0.00	0
Creative Industries	1.00	0.43	0.60	7
Energy	0.75	0.50	0.60	6
High tech systems and materials	0.67	0.40	0.50	10
Life sciences and health	0.00	0.00	0.00	0
Transportation and storage	0.50	0.33	0.40	3
Horticulture and raw materials	1.00	0.40	0.57	5
Water	1.00	0.33	0.50	9
Average / Total	0.85	0.41	0.55	41

Table 16 classification enterprises based on NACE code

Accuracy on the test set: 0.46

Top-sector	Precision	Recall	F1-score	Support
Other	0.17	0.10	0.12	10
Agriculture	0.78	0.54	0.64	46
Chemistry	0.64	0.35	0.45	20
Creative Industries	0.85	0.72	0.78	40
Energy	0.60	0.75	0.67	8
High tech systems and materials	0.84	0.57	0.68	54
Life sciences and health	1.00	0.68	0.81	19
Transportation and storage	0.88	0.37	0.52	19
Horticulture and raw materials	0.62	0.38	0.48	13
Water	0.79	0.69	0.73	16
Average / Total	0.78	0.55	0.64	245

3.5.3 Website complexity

We compared the situation of using only the words from the homepage to feed into the classifier (Table 17) , with the situation where the words come from both the homepage and one layer underneath (Table 11). Besides an indicator for complexity this experiment is also important for efficiency as scraping solely the homepage is more time efficient. The results, shown in Table 17, illustrated that the homepage contained sufficient information to make a prediction. Scraping additional information from other pages is thus not necessary.

We first compare the average performance at top-sector level of both situations. For the 'homepage only' situation we then found a slightly lower precision and a slightly higher recall that for the ‘homepage plus one layer’ situation, whereas the F1 score

was nearly the same. This implies that for the average performance, the complexity of the website is not a decisive factor.

Table 17 Classification report when predicting with 'homepage only'

Accuracy on the test set: 0.49

Top-sector	Precision	Recall	F1-score	Support
Other	0.23	0.30	0.26	10
Agriculture	0.78	0.66	0.71	47
Chemistry	0.69	0.55	0.61	20
Creative Industries	0.74	0.68	0.71	47
Energy	0.68	0.93	0.79	14
High tech systems and materials	0.80	0.52	0.63	64
Life sciences and health	0.76	0.68	0.72	19
Transportation and storage	0.83	0.45	0.59	22
Horticulture and raw materials	0.56	0.56	0.56	18
Water	0.52	0.52	0.52	25
Average / Total	0.71	0.59	0.64	286

Table 18 Classification report when predicting with 'homepage plus one layer'

Accuracy on the test set: 0.45

Top-sector	Precision	Recall	F1-score	Support
Other	0.17	0.10	0.12	10
Agriculture	0.79	0.55	0.65	47
Chemistry	0.64	0.35	0.45	20
Creative Industries	0.86	0.68	0.76	47
Energy	0.64	0.64	0.64	14
High tech systems and materials	0.81	0.55	0.65	64
Life sciences and health	1.00	0.68	0.81	19
Transportation and storage	0.80	0.36	0.50	22
Horticulture and raw materials	0.70	0.39	0.50	18
Water	0.82	0.56	0.67	25
Average / Total	0.78	0.53	0.63	286

Next, we compared the performance differences between the two situations for the different top-sectors. For two of the top-sectors, namely ´high tech systems and materials´ and ´Transportation and storage´ the performance in both situations was similar. For four of the top-sectors the precision was clearly lower for the ´homepage only´ situation, namely for ´Creative Industries´, ´Life sciences and health´, ´Horticulture and raw materials´ and ´Water´. For three top-sectors the recall was clearly higher for the ´homepage only´ situation, namely for ´Agriculture´, ´Chemistry´ and for ´Energy´. Finally, for the additional category ´other´ the recall was smaller for the ´homepage only´ situation compared to the ´homepage plus one layer´ situation.

We thus found that the performance at top-sector level varied between the two situations, where the ´homepage plus one layer´ was not always better than the ´homepage only´ situation.

4. Discussion

The objective of this study was to evaluate the suitability of text mining techniques to automatically classify enterprises to a standard classification of economic activity. As an example of a standard classification, we used the top-sector classification. This was done by studying the effect of: different weighting methods, single-label versus multi-label prediction, different methods to select a dictionary and different classifiers. Furthermore the study evaluated three possible complexity influencing characteristics: the size of the enterprise, whether an enterprise was labelled on the basis of the NACE code or by membership of a trade organisation and the number of webpages (within a website) on which a prediction was made.

The results showed that overall adding an Inverse Document Frequency (IDF) weighting to the Term Frequency (TF) sometimes does and sometimes does not lead to a performance increase. Although TF-IDF has proven to be a good pre-processing step for information retrieval with text (Pazzani et al., 1996), this was not always the case in the present study. Part of this effect may be because the input for which classifiers perform best differs between classifiers (Weiss et al., 2010). Based on the current study, we advise to compute both the results based on the TF and on TF-IDF when classifying economic activity, and then select the best result.

In our study we compared a multi-label, corresponding to the real situation, with a single-label analysis which was in fact a simplification. Because there are enough classifiers that can handle a multi-label classification (Tsoumakas and Katakis, 2006), the multi-label approach is preferred. We found that recall and accuracy were better for the single-label approach whereas the precision was better with the multi-label approach. An explanation might be that the single-label approach is forced to make a decision also for each instances where there classes overlap. The multi-label algorithm in a 'one versus the rest' set-up is more conservative in assigning labels. When a label is predicted in the multi-label approach it is more precise, but overall it predicts not all the labels.

The preferred dictionary set depended on the classifier. For the NB classifier, best performance was found with the pooled dictionary: pooling knowledge-based with the automatic feature selection. For the other classifiers (Decision tree, RF, SVM and LR) best performance was found with the intersection dictionary. The overlap between the two dictionaries was relatively small, only 8% of the complete set of words was in the overlap. This result indicates that the NB can better handle high dimensional features in combination with a limited number of training instances. The other algorithms need shorter features, that contain the terms with the most information, to tackle the curse of dimensionality (Indyk and Motwani, 1998).

The performance of the knowledge-based dictionary set was just slightly better than an automatically generated feature set of the same size. This result was unexpected since these knowledge-based dictionaries have a good track record in sentiment

analysis (Kouloumpis et al., 2011). For future research, it is worthwhile to invest more time in the feature selection. Other studies indicated that most time and effort is invested in pre-processing while having good features is very important for the outcome (Lohr, 2014). It may be necessary to spend more time and effort in developing a knowledge-based set specific for the automatic coding of economic activities with machine learning algorithms. The necessary terms may differ from the terms human coders prefer.

A quite surprising result from the classifier comparison is the good performance of the Naïve Bayes classifier. Earlier studies showed that Naïve Bayes often is used as a benchmark classifier and the SVM is the best algorithm for text mining (Rennie et al., 2003; Aggarwal and Zhai, 2012). An explanation for our deviation result may be that the Naïve Bayes classifier can deal very well with a limited number of training examples. In our study, the sample size was relatively small. SVM is potentially the best performing classifier but it requires sufficient samples (Hotho et al., 2005; Aggarwal and Zhai, 2012). Still, the results are unexpected since SVM and NB both can be rewritten into the same linear form(Weiss et al., 2010), where in SVM weights are added to the expression. We cannot completely rule out that we would have could have improved the SVM results with a broader grid search.

We also experimented with applying an ensemble learning algorithm. We applied just a simple algorithm, which resulted in only a slight improvement over the basic classifier. More advanced ensemble methods can probably produce better outcomes. This is supported by other studies that showed that using 'Adaboost' or a 'weighted voting classifier' can be good way to increase the performance on text classification (Sebastiani, 2002; Witten et al., 2016, pp. 479-501; Silla and Freitas, 2011).

Final result to be discussed concerns the evaluation of complicating characteristics. The results showed that the developed method is slightly better in classifying larger enterprises than one-man enterprises. A larger size of the test set is needed before we can be sure about this result. Our impression is that larger enterprises have more advanced websites with more information on them. Whether larger enterprises are also easier to classify for other classifications is an interesting topic for further research. With regard to the origin of the labels (derived from NACE codes or from trade organization membership) we found that the size of the test set was too small to draw clear conclusions. Finally, our website complexity comparison showed that, averaged over the top-sectors, information on the homepage was a sufficient predictor for the top-sector classification compared the use of 'homepage plus one layer'. However, at top-sector level the performance for the 'homepage plus one layer' was not always as good as the performance for the 'homepage only' situation. In future, we aim to find out what caused this result and how we can improve this.

We found that the top-sector classification was an interesting case study to get an idea of the potential of text mining to predict economic activity. Prediction of 10 classes was already shown to be difficult with the use of a classifier and 30 classes was even more difficult. Using the top-sector case study as an example to predict economic activity also had its limitations. A first limitation is that top-sector is a

multi-label classification whereas prediction of main activity is single-label. On the other hand: if one is interested to predict a multiple economic activities (main and side activities) per enterprise then the problem would also be multi-label. A second limitation is that some of the categories contain a rather heterogeneous set of economic activities which makes it difficult to predict, especially given the limited size of the test set. Prediction of economic activity in terms of NACE code classes may be more successful since those classes are more homogeneous in activity.

There are a number of issues that are interesting to investigate in order to improve the results of this case study and in automatically classifying economic activity. A first set of issues concerns the features that are used in the machine learning algorithms:

- It would be useful to compare different automatic feature selection methods. Since the full set of words contains about 240.000 words, this requires sufficient computing capacity;
- A closer look into those categories and websites that cannot be predicted well and trying to find its causes might be useful, That may form the basis to find improvements, for instance in terms of the kinds of features used;
- One might improve on the knowledge-based dictionary. It would be interesting to study whether class-specific dictionaries give an improved classifier performance;
- One might add additional features to predict the categories such as web site language, n-grams, context and position of the words, total number of words and so on;
- One might compare different kinds of website content extraction methods (Sozzi, 2017).

A second set of issues concerns improvement of the machine learning algorithms:

- The performance of the machine learning algorithms might improve when we use larger training-validation set. We can validate this by constructing a learning curve;
- We are aware that the NACE code attribute in the GBR that we used contains measurement errors, especially for the smaller enterprises, see for instance (Van Delden et al., 2016). The performance of the machine learning algorithm might profit from a manual verification of the NACE codes in the training-validation set. We did not do this so far, because it is a very time-consuming and 'specialist' activity;
- It might be interesting to investigate whether prediction of sub-sector level might be improved by using a technique that combines prediction at two different aggregate levels, such as presented in Gweon et al. (2017);
- In the current paper we used a one-versus the rest approach for the multi-class prediction. Two other known methods are the use of binary classifications for each pair of categories and the use of a tree of binary classifications. Those other methods might give an improvement;
- We experimented with the use of a very simple ensemble method. It would be useful to analyse whether other ensemble method give better results. For instance, when we use a larger the training-validation set, we could apply bagging;

- Another type of future research that might be interesting is combining text mining and image recognition. For instance, agricultural activities on a website may be recognized by the picture shown on the webpages.

All in all, in the present study some important steps have been taken in developing a method to automatically classify economic activity but before full scale implementation is possible some further steps need to be taken.

5. Conclusions

Based on the top-sector case study we investigated the effect of a number of factors for a text mining classifier to be successful. Concerning TF versus TF-IDF word weighting we found that neither approach performed best under all conditions and that it is best to include that factor in the grid search to optimise the parameters. We further found that recall and accuracy were better for the single-label approach whereas the precision was better with the multi-label approach. Furthermore, the best performance was found for the Naïve Bayes estimator using a pooled set of an automatic feature selection with a knowledge-based feature dictionary, for both top-sector and sub-sector prediction. For the sub-sector classification, Random Forest, Logistic Regression and SVM in combination with the intersection dictionary was shown to be second best compared to Naïve Bayes.

Furthermore, we evaluated the effect of complexity characteristics on classifier performance. Results indicated that the top-sector categories one-man enterprises are more difficult to predict than those of larger enterprises. Furthermore, we found that predictions solely based on words from the homepage were close to those base on homepage and one underlying webpage layer. The sample size was too small to draw conclusions whether prediction performance varies with label origin (NACE code versus trade organisation membership).

Our research question was: “Are text mining techniques suitable to automatically classify enterprises to a standard classification of economic activity?”. As a benchmark for full automatic classification, we used Figure 3 in Gweon et al. (2017), where their poorest machine learning method yielded an accuracy of just below 53% while their best method yielded a 65% accuracy , for a large number of occupation classes. We arrived at an accuracy of 51,2% (Table 12) for our best results, on 10 top-sector classes. That implies there is a challenge to improve on the accuracy. We have already mentioned a number of points for improvement (see discussion section). While the precision (80%) of the best performing method is sufficient, the accuracy and recall are not good enough. Furthermore, the results to predict sub-sector are clearly less good than those for top-sectors. Directions to improve the results are: use of more advanced ensemble methods, combining predictions at different aggregate levels and improvement of the features used in the classifiers.

6. Appendix

Table 19 Evaluating the effect of TF-IDF weighting for predicting sub-sectors

Classifier	Variation	Parameters	Validation score (F1)	A	P	R	F1
KNN	TF	n_neighbours = 3 metric = Cosine	0.335	0.228	0.56	0.26	0.32
	TF-IDF	n_neighbours = 3 metric = Cosine	0.414	0.290	0.60	0.32	0.39
RF	TF	Criterion = 'gini' min_samples_split = 1 n_estimators = 5	0.263	0.167	0.52	0.20	0.28
	TF-IDF	Criterion = 'gini' min_samples_split = 1 n_estimators = 5	0.278	0.167	0.49	0.21	0.29
NB	TF	Alpha = 0.0001 fit_prior = False	0.427	0.282	0.39	0.59	0.46
	TF-IDF	Alpha = 0.001 fit_prior = True	0.434	0.253	0.56	0.36	0.42
SVM	TF	C = 10 Gamma = 0.1 Kernel = 'linear'	0.354	0.215	0.53	0.27	0.34
	TF-IDF	C = 8 Gamma = 0.001 Kernel = 'linear'	0.395	0.278	0.56	0.36	0.42
LR	TF	Alpha = 0.0001 n_iter = 100 penalty = ' elasticnet'	0.379	0.242	0.53	0.32	0.39
	TF-IDF	Alpha = 0.00001 n_iter = 100 penalty = 'elasticnet'	0.402	0.282	0.53	0.38	0.43

Table 20 Jaccard index NACE dictionary filter for sub-sector[3]and top-sector

	Ag_wh	Ag_or	Ag_pc	Ag_ma	Ch_pe	Ch_ch	Ch_pl	Cr_cr	Cr_cu	Cr_ar	Cr_me	En_pe	En_re	Hi_me	Hi_ot	Hi_ma	Hi_tr	Li_ph	Li_me	Li_re	Tr_wa	Tr_tr	Ho_or	Ho_pr	Wa_co	Wa_bo	Wa_su
Ag_wh	1																										
Ag_or	0.14	1																									
Ag_pc	0.21	0.15	1																								
Ag_ma	0.34	0.01	0.13	1																							
Ch_pe	0.01	0.03	0.01	0.02	1																						
Ch_ch	0.01	0.10	0.0	0.02	0.02	1																					
Ch_pl	0.01	0.01	0.01	0.01	0.02	0.01	1																				
Cr_cr	0.01	0.02	0.03	0.01	0.03	0.01	0.01	1																			
Cr_cu	0.01	0.0	0.01	0.0	0.0	0.0	0.0	0.01	1																		
Cr_ar	0.02	0.02	0.03	0.01	0.03	0.01	0.01	0.05	0.05	1																	
Cr_me	0.02	0.02	0.03	0.01	0.03	0.01	0.1	0.04	0.04	0.12	1																
En_pe	0.01	0.02	0.02	0.01	0.10	0.02	0.01	0.04	0.0	0.07	0.05	1															
En_re	0.02	0.02	0.02	0.02	0.09	0.02	0.02	0.03	0.0	0.03	0.05	0.15	1														
Hi_me	0.01	0.01	0.01	0.01	0.01	0.02	0.13	0.01	0.0	0.01	0.0	0.02	0.01	1													
Hi_ot	0.01	0.03	0.02	0.01	0.02	0.01	0.01	0.06	0.0	0.03	0.03	0.04	0.02	0.03	1												
Hi_ma	0.07	0.12	0.06	0.01	0.01	0.01	0.04	0.01	0.0	0.01	0.01	0.02	0.02	0.06	0.08	1											
Hi_tr	0.01	0.01	0.01	0.01	0.02	0.05	0.34	0.02	0.0	0.04	0.02	0.03	0.04	0.07	0.05	0.04	1										
Li_ph	0.01	0.03	0.01	0.01	0.10	0.02	0.02	0.03	0	0.03	0.03	0.07	0.08	0.01	0.02	0.01	0.02	1									
Li_me	0.01	0.02	0.01	0.01	0.04	0.01	0.01	0.03	0	0.05	0.02	0.07	0.04	0.02	0.02	0.10	0.04	0.05	1								
Li_re	0	0.05	0	0	0	0	0	0.0	0	0	0.0	0	0	0.0	0.02	0.01	0.0	0.10	0.12	1							
Tr_wa	0.02	0.0	0.02	0.1	0.0	0.0	0.0	0.02	0.02	0.03	0.03	0.03	0.02	0.00	0.05	0.01	0.06	0	0	0	1						
Tr_tr	0.02	0.0	0.03	0.01	0.02	0.0	0.0	0.01	0.03	0.02	0.03	0.04	0.01	0.01	0.10	0.0	0.03	0	0	0	0.19	1					
Ho_or	0.12	0.0	0.13	0.15	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	0	0	0.01	0.0	1				
Ho_pr	0.11	0.0	0.12	0.11	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.0	0.0	0.00	0.0	0.0	0	0	0.0	0.01	0.46	1			
Wa_co	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.0	0.01	0.03	0.06	0.02	0.02	0.0	0.02	0	0.0	0.01	0.0	0.0	1		
Wa_bo	0.01	0.02	0.01	0.1	0.04	0.02	0.02	0.03	0.0	0.07	0.03	0.10	0.05	0.03	0.16	0.02	0.07	0.04	0.08	0	0.02	0.23	0.0	0.01	0.06	1	
Wa_su	0.0	0.0	0.01	0.0	0	0	0	0	0	0	0.01	0.08	0.11	0.0	0.01	0.0	0.0	0	0	0	0.0	0.01	0.0	0.01	0.0	0.01	1

	Agriculture	Chemistry	Creative industry	Energy	High tech	Life science	Transportation	Horticulture	Water
Agriculture	1								
Chemistry	0.034	1							
Creative industry	0.025	0.011	1						
Energy	0.020	0.017	0.038	1					
High tech	0.058	0.061	0.027	0.017	1				
Life science	0.014	0.011	0.026	0.050	0.061	1			
Transportation	0.027	0.007	0.024	0.032	0.036	0.0	1		
Horticulture	0.116	0.003	0.004	0.040	0.001	0.001	0.011	1	
Water	0.008	0.009	0.033	0.120	0.056	0.056	0.117	0.006	1

[3] Row and columns names abbreviation of sub sector names. First two letters are the first letters of the top sector, the last two letter refer to sub sector specification

Table 21 Evaluating the effect of the dictionaries for predicting sub-sectors

Classifier	Variation	Parameters	Validation score (F1)	A	P	R	F1
KNN	NACE	Use_idf = True n_neighbours = 3 metric = Cosine	0.294	0.149	0.40	0.16	0.22
	K-best	Use_idf = True n_neighbours = 3 metric = Cosine	0.414	0.290	0.60	0.32	0.39
	Pool	Use_idf = True n_neighbours = 3 metric = Cosine	0.355	0.187	0.50	0.22	0.29
	Intersect	Use_idf = True n_neighbours = 3 metric = Cosine	0.301	0.198	0.48	0.23	0.29
RF	NACE	Use_idf = False Criterion = 'gini' min_samples_split = 5 n_estimators = 5	0.209	0.066	0.36	0.10	0.15
	K-best	Use_idf = False Criterion = 'gini' min_samples_split = 5 n_estimators = 5	0.189	0.095	0.44	0.13	0.19
	Pool	Use_idf = True Criterion = 'gini' min_samples_split = 1 n_estimators = 5	0.248	0.116	0.53	0.15	0.22
	Intersect	Use_idf = True Criterion = 'gini' min_samples_split = 1 n_estimators = 5	0.259	0.179	0.56	0.21	0.33
NB	NACE	Use_idf = True Alpha = 0.00001 fit_prior = False	0.409	0.212	0.67	0.24	0.34
	K-best	Use_idf = False Alpha = 0.0001 fit_prior = False	0.387	0.270	0.42	0.50	0.44
	Pool	Use_idf = True Alpha = 0.001 fit_prior = True	0.434	0.253	0.55	0.36	0.42
	Intersect	Use_idf = True Alpha = 0.00001 fit_prior = False	0.359	0.222	0.35	0.60	0.42

Table 20 (cont.)

Classifier	Variation	Parameters	Validation score (F1)	A	P	R	F1
SVM	NACE	Use_idf = True C = 10 Gamma = 0.01 Kernel = 'linear’	0.392	0.253	0.50	0.30	0.36
	K-best	Use_idf = False C = 10 Gamma = 0.01 Kernel = 'linear'	0.402	0.253	0.61	0.31	0.39
	Pool	Use_idf = True C = 8 Gamma = 0.1 Kernel = 'linear'	0.398	0.228	0.57	0.29	0.37
	Intersect	Use_idf = True C = 8 Gamma = 0.001 Kernel = 'linear'	0.395	0.277	0.56	0.36	0.42
LR	NACE	Use_idf = True Alpha = 0.000001 n_iter = 10 penalty = 'elasticnet'	0.378	0.249	0.41	0.34	0.35
	K-best	Use_idf = True Alpha = 0.00001 n_iter = 10 penalty = 'l2'	0.401	0.257	0.57	0.32	0.39
	Pool	Use_idf = True Alpha = 0.000001 n_iter = 10 penalty = ‘l2’	0.392	0.232	0.52	0.30	0.37
	Intersect	Use_idf = True Alpha = 0.00001 n_iter = 100 penalty = 'l2'	0.401	0.277	0.52	0.37	0.42

Table 22 Classification report Voting Classifier, intersection dictionary for predicting top-sectors

Accuracy on the test set: 0.51

Top-sector	Precision	Recall	F1-score	Support
Other	0.00	0.00	0.00	9
Agriculture	0.71	0.52	0.60	46
Chemistry	1.00	0.37	0.54	19
Creative Industries	0.86	0.79	0.82	47
Energy	0.88	0.68	0.77	22
High tech systems and materials	0.81	0.60	0.69	65
Life sciences and health	0.87	0.87	0.87	15
Transportation and storage	1.00	0.43	0.61	23
Horticulture and raw materials	0.75	0.44	0.56	27
Water	0.75	0.60	0.67	25
Average / Total	0.80	0.58	0.66	298

Table 23 Classification report NB classifier, K-best dictionary for predicting sub-sectors

Accuracy on the test set: 0.27

Top-sector – sub-sector	P	R	F1	Support
Agriculture - Wholesale and retail Trade	0.46	0.75	0.57	8
Agriculture - Primary production	0.29	0.31	0.30	16
Agriculture - Manufacture of food products	0.45	0.83	0.59	12
Agriculture - Other	0.67	0.40	0.50	10
Chemistry - Manufacture of refined petroleum products	0.00	0.00	0.00	0
Chemistry - Chemical industry	0.00	0.00	0.00	6
Chemistry - Manufacture of rubber and plastic products	0.50	0.57	0.53	14
Creative Industries - Creative services	0.32	0.43	0.36	14
Creative Industries - Cultural heritage	0.52	0.92	0.67	13
Creative Industries – Art	0.29	0.17	0.21	12
Creative Industries - Media and entertainment industry	0.25	0.38	0.30	8
Energy - Extraction of crude petroleum and gas	0.50	0.67	0.57	3
Energy - Sustainable energy	0.18	0.50	0.26	6
Energy - Related activities	0.43	0.75	0.55	4
High tech - Manufacture of metal products	0.40	0.36	0.38	11
High tech - Manufacture of machinery	0.59	0.65	0.62	20
High tech - Manufacture of transport equipment	0.50	0.43	0.46	14
High tech – Other	0.27	0.33	0.30	9
Life sciences – Pharmaceutical	0.20	0.17	0.18	6
Life sciences - Manufacture of medical instruments	0.67	1.00	0.80	10
Life sciences - Research and development	0.14	0.33	0.20	3
Transportation – Transport	0.38	0.62	0.48	8
Transportation - Warehousing and support activities	0.67	0.43	0.52	14
Horticulture - Primary production	0.33	0.62	0.43	8
Horticulture – Other	0.23	0.30	0.26	10
Water - Construction of water projects	0.35	0.78	0.48	9
Water - Building and repairing of ships and boats	0.14	0.33	0.20	3
Water - Water collection, treatment and supply	1.00	0.33	0.50	6
Water – Consultancy	0.62	0.71	0.67	7
Other	0.17	0.20	0.18	10
Average / Total	0.42	0.50	0.44	274

7. References

Aggarwal, C. C. and Zhai, C. (Eds.). (2012). Mining text data. *Springer Science and Business Media*.

Aphinyanaphongs, Y., Fu, L. D., Li, Z., Peskin, E. R., Efstathiadis, E., Aliferis, C. F., Baxter, P. and Jack, S. (2008). Qualitative case study methodology: Study design and implementation for novice researchers. *The qualitative report*, 13(4), 544-559.

Buelens, B., Daas, P., Burger, J., Puts, M. and Van den Brakel, J. (2014). Selectivity of Big data. CBS report, February 2014. This paper presented at the Statistics Netherlands Advisory Council meeting on 11 Feb. 2014.

CBS (2016). Monitor topsectoren 2016. Methodebeschrijving en tabellenset. Den Haag: Centraal Bureau voor de Statistiek. (in Dutch)

Chen, B-C.., Creecy, R.H. and Appel, M.V. (1993). On Error Control of Automated Industry and Occupation Coding. *Journal of Official Statistics* 9(5), 729–745 (1993)

Cheung, P. (2012). Big Data, Official Statistics and Social Science Research: Emerging Data Challenges*. presentation at the World Bank*.

Daas, P. (2012). Big Data and official statistics. Sharing Advisory Board: Software Sharing Newsletter, 7, 2-3.

Daas, P. J., Puts, M. J., Buelens, B., and Van den Hurk, P. A. (2015). Big data as a source for official statistics. *Journal of Official Statistics*, 31(2), 249-262.

Delden, A. van, Scholtus, S. and J. Burger (2016). Accuracy of mixed-source statistics as affected by classification errors. *Journal of Official Statistics*, 32(3), 619–642.

Dumais, S., Platt, J., Heckerman, D., and Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management* (ACM-CIKM) (pp. 148-155).

El-Halees, A. M. (2015). Arabic text classification using maximum entropy. *IUG Journal of Natural Studies*, 15(1).

Griffioen, R., de Haan, J., and Willenborg, L. (2014, May). Collecting clothing data from the Internet. In *Proceedings of Meeting of the Group of Experts on Consumer Price Indexes* (pp. 26-28).

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M. and Steiner, S. (2017). Three methods for occupation coding based on statistical learning. *Journal of Official Statistics*, 33(1), 101-122.

Hackl, P. (2016). Big Data: What can official statistics expect? *Statistical Journal of the IAOS*. 32.1: 43-52.

Hassani, H., Saporta, G. and Silva, E. S. (2014). Data mining and official statistics: the past, the present and the future. *Big Data*, 2(1), 34-43.

Hearst, M. (2003). What is text mining. SIMS, UC Berkeley.

Hotho, A., Nürnberger, A. and Paaß, G. (2005). A brief survey of text mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology* 20 (1), 19-62.

Indyk, P. and Motwani, R. (1998). Approximate nearest neighbours: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing* (pp. 604-613).

Jung, Y., Yoo, J., Myaeng, S. H. and Han, D. C. (2008, September). A web-based automated system for industry and occupation coding. In *proceedings of the International Conference on Web Information Systems Engineering* (pp. 443-457). Berlin: Springer.

Kouloumpis, E., Wilson, T. and Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg!. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011* (pp. 538-541). AAAI Press.

Lohr, S. (2014), "For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights". *New York Times*, 17 August 2014.

Pazzani, M. J., Muramatsu, J. and Billsus, D. (1996). Syskill & Webert: Identifying interesting web sites. Proceedings of the National Conference on Artificial Intelligence, Portland, OR, Vol. 1 (pp. 54-61).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830.

Porter, M. F. (2001). Snowball: A language for stemming algorithms. Accessed at http://snowball.tartarus.org/texts/introduction.html

Reimsbach-Kounatze, C. (2015), "The Proliferation of "Big Data" and Implications for Official Statistics and Statistical Agencies: A Preliminary Analysis", *OECD Digital Economy Papers*, No. 245, OECD Publishing, Paris.

Rennie, J. D., Shih, L., Teevan, J. and Karger, D. R. (2003, August). Tackling the poor assumptions of naive bayes text classifiers. *Proceedings of the international conference on machine learning* (*ICML)* 3, pp. 616-623).

Rigter, I. (2017, 02 may). The power of big data: measuring the internet economy. Retrieved from https://blog.dataprovider.com/the-power-of-big-data-measuring-internet-economy/.

Roelands, M.J. (2017). Classifying economic activity with business websites using text-mining. Master thesis for Tilburg University, July 2017.

Sebastiani, F. (2002) Machine learning in automated text categorization. *ACM Computing Surveys* 34(1), 1–47.

Silla, C. N. and Freitas, A. A. (2011). A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1-2), 31-72.

Sozzi, A. (2017). Measuring Sustainability Reporting using Web Scraping and Natural Language Processing. Office of National Statistics. Available at https://github.com/AlessandraSozzi/MSc-dissertation (accessed 30-10-2017).

Stockwell, D. R. and Peterson, A. T. (2002). Effects of sample size on accuracy of species distribution models. *Ecological modelling*, 148(1), 1-13.

Struijs, P., Braaksma, B. and Daas, P. J. (2014). Official statistics and big data. *Big Data and Society*, 1(1).

Tam, S. M. and Clarke, F. (2015). Big data, official statistics and some initiatives by the Australian Bureau of Statistics. *International Statistical Review*, 83(3), 436-448.

Tarnow-Mordi, R. (2017). The intelligent coder: developing a machine-learning classification system. Article in ABS Methodological News 2017. Available at http://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/1504.0Main%20Features5Sep%202017?opendocument&tabname=Summary&prodno=1504.0&issue=Sep%202017&num=&view=

Thompson, M., Kornbau, M. E. and Vesely, J. (2012). Creating an automated industry and occupation coding process for the American Community Survey. US census Bureau.

Tsoumakas, G. and Katakis, I. (2006). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3).

Van Asch, V. (2013). Macro-and micro-averaged evaluation measures.

Weiss, S. M., Indurkhya, N., Zhang, T. and Damerau, F. (2010). Text mining: predictive methods for analyzing unstructured information. *Springer Science and Business Media*.

Williams, N. (2006). ACTR/IDBR Test Evaluation Report. In Office for National Statistics, *Survey methodology bulletin*. No. 57, pp. 33-38.

Witten, I. H., Frank, E., Hall, M. A. and Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Zhang W, Yoshida T and Tang X (2011). A comparative study of TF*IDF, LSI and multi-words for text classification. *Expert Systems with Applications* 38(3): 2758-2765.

Acknowledgements

We thank Tommy Span and Remco Kaashoek for providing the case study materials and their time to explain the details of the case study. Furthermore, we thank Ali Hürriyetoglu for his useful comments on earlier versions of the paper.

Colofon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
http://www.cbs.nl

Prepress
Statistics Netherlands, Studio BCO

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbs.nl/information