IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Web Data Sources and Processing

Dr.K.Vijayan
Department of Information Technology
CMS College of Science And Commerce
E-Mail : vijayan_709@yahoo.co.in

Abstract

Statistics Netherlands (CBS) is interested in using Natural Language Processing (NLP) to classify companies that are not included in Community Innovation Survey (CIS) to obtain reliable data with regard to the location of innovation activities. Various machine learning methods were applied with favorable results in the past. In recent years, growing attention is paid to combining predictions of multiple models. ensemble approach is investigated in relation to predicting innovative companies based on their website text. It was found that the stacking algorithm provided the best accuracy out of all the models but with considerable training time. Depending on the seed and a random selection of the training data, the stacking algorithm provided an improvement of up to 1%. Other ensemble algorithms presented did not improve the accuracy compared to the best-performing individual model.
The main point of this chapter is to cover the data source and webs craping process. Moreover, preprocessing steps and mathematical representation of words extracted from the websites will be discussed.

1.1 Data sources

The primary source of data is the 2016 CIS survey. The survey covered theperiod of 2014-2016 and included all companies with 200 or more employees and a stratified sample of companies with 10 or more employees. Stratified sample is a method of sampling where the population is divided into several subpopulations (strata) that are based on the size and type of economic activityof companies (Commission and Eurostat, 2008). how a stratified sampling is performed. In the case of the Netherlands, in 2018 CIS survey that is very similar to the 2016 one, there are 182 strata consisting of 6 size classes and 41 NACE economic activity classes such as education, construction, telecommunication, etc. (Eurostat and Netherlands, 2023). From each group, separate samples are drawn and the number of those samples depends on the number of units in each stratum, the size of the companies, and the variability of the main variables within each subpopulation such as non-response rate, location, industry type, etc. (OECD and Eurostat, 2018). In total 10000 companies are chosen to be in the 2016 survey. Thereafter, an attempt was o the companies included in the survey. 6214 companies were selected based on link availability and scraping possibility since not all companies with a website allow scraping. Subsequently, the resulting dataset consists of 3,338 innovative companies and 2,876 non-innovative.

One of the major issues in the previous study from Daas and vander Doef (2020) was the concept drift. This m that the classification power of the model deteriorated over time because of changes in websites and companies t self. As a response, in the follow-up study by Daas and Jansen (2020)it was attempted to include more sam in the training data. The websites

of companies included in the Dutch Business Register were classified by the previous model. Labels

obtained this way were added to the dataset and websites were scraped again one year later. The dataset obtained in this way consisted of 25618 records. A new classification was performed and the resultswere promising with an accuracy of 88% from the new Logistic Regression model. The same dataset was used by Peer boom (2023) to validate a concept- drift resilient approach of using the SBERT model for a vector representation ofthe data. Because concept drift is an important issue encountered in innovationresearch, the same dataset will be used in this thesis. This thesis will not discuss the concept drift per se but can serve as a basis for the future research in that regard. 20000 companies will be randomly sampled to train and test all the models presented in this thesis. The dataset is fairly balanced with about 40% of the companies being innovative.

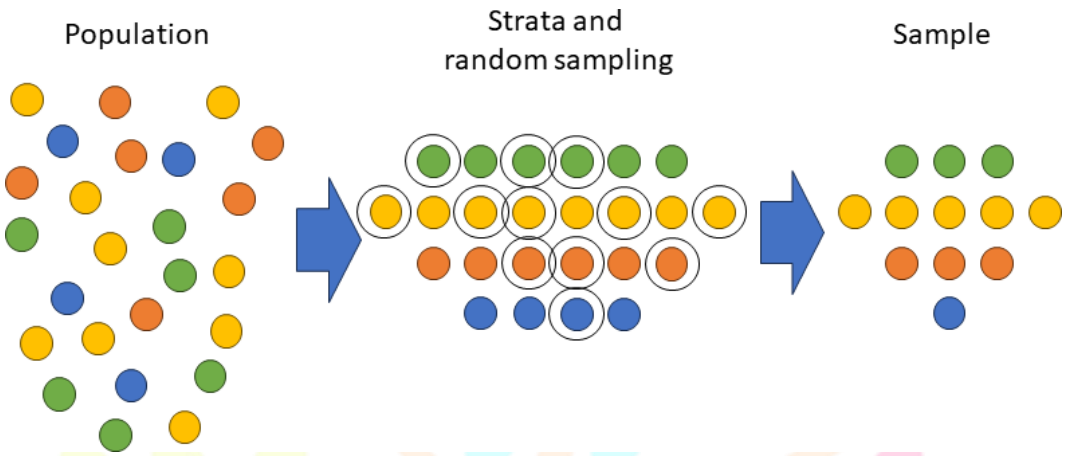


Figure 1.1: Stratified Sample

1.2 Pre-processing of the website text

The first step of every text analysis is identifying the language of the document. The later employed pre-processing like lemmatization and stop-word removal are language-dependent. *Lang detect* library is used to determine the language of each website. Whenever a non-Dutch website was found, it was treated as an English website.

Secondly, for better accuracy tokenization is performed. Tokenization is beneficial for better analysis results as every word forms a token that latercan be adjusted or removed to extract maximum information value out ofthe document (Chai, 2023). In order to divide the text corpus into tokens correctly, all numbers and punctuation marks are converted to spaces, and multiple spaces are removed by using regular expressions. Regular expressions are an efficient and simple way of text manipulation (Friedl, 2006). In this way, missing spaces after commas and dots that might occur on the websites are accounted for. Furthermore, all letters are converted into lowercase.

Thirdly, to obtain the best value out of tokens the most repeating wordsare removed that do not hold much meaning for the computer, and change the remaining words such that the text corpus is smaller. Stop words are widely used phrases and conjunctions that typically do not hold much information for Natural Language Processing algorithms. A package *nltk* is used to delete stop words based on publicly available lists. Moreover, short abbreviations like 'de', 'nl', etc. are usually part of domains hence removing them is recommended despite the potential increase in model performance that follows from the inclusion (Daas and van der Doef, 2020). A website full of domain names should not be classified as innovative. Consequently, it is decided to remove all words with 2 or fewer characters.

Lastly, decreasing the number of words and preventing data sparsity stemming or lemmatization is considered by many researchers (Khyani and B S, 2021). Stemming is reducing a word to its root form known as word stem by removing a certain number of characters at the end or changing them depending on the stemming algorithm used. Similarly, lemmatization occurs when a word is converted into its grammatically

base form called lemma. The difference between the two is presented in the following table for English examples:

Table 1.2 Stemming and Lemmatization Daas and van der Doef (2020) used stemming in their work and achieved very good results hence the same will be applied in this thesis.

Word	Stemming (Lancaster) (spacy)	Lemmatization
is	is	be
better	bet	well
worse	wors	bad

1.3 Vector representations

Before any (machine learning) algorithms are applied, there is a need fora numerical representation of the data. Term Frequency-Inverse Document Frequency (tf-idf) could be created that would provide a vector of weight numbers. Each weight would correspond to a combination of frequency in which the word appears in the website itself as well as all of the websitesin the data set. Term frequency and inverse document frequencies are two variables that are used to compute the score. It is a popular method and despite being used for a long time, it still remains favored in many domains, for example in research-paper recommender systems (Beel et al., 2016). In order to compute tf-idf frequency the following variables and sets are used:

Variables and Sets

w : Word
d : Document (website)
W : Set of all words in all of the documents
D : Set of all documents
$|D|$: Number of documents
$|W|$: Number of

words in all of the documents
$f_{d,w}$: Number of occurrences of word w in document d
n_w: Number of occurrences of word w in all documents Matrix of tf-idf scores y : Target variable representing the binary outcomeof the survey
The tf-idf start with computation of tf parameter:$f_{d,w}$
$tf(d, w)$ is calculated by dividing number of

occurrences of word w in document d by the total number of terms in document d via

formula:$idf(D, w) = \log \frac{|D|}{n_w}$

$idf(D, w)$ measures how common word w is across all documents. The useof the logarithm serves as smoothing the frequencies of words. Without thelogarithm, words at the end of frequency spectrum would be indistinguishable from one another in terms of their idf score. Logarithmic scale effectively linearizes the scores. The final tf-idf equation presents as follows:

$tf\text{-}idf(|D|, w) = tf(d, w) \cdot idf(|D|, w)$

$tf\text{-}idf(d, w, D)$ is the final score of word w to measure its rarity among allthe words in the document. A higher score implies that a word occurs lessfrequently. Using $tf\text{-}idf(d, w, D)$ formula, every entry of matrix X, $X \in \mathbb{R}^{|D| \times |W|}$ is defined as:

$$x_{dw} = tf\text{-}idf(|D|, w) \qquad \forall d \in D, \forall w \in W$$

Finally, target variable $y \in \{0, 1\}^{|D|}$ is a binary vector where:

$$d = \begin{cases} 0 & \text{if a company is not innovative} \\ 1 & \text{if a company is innovative} \end{cases}$$

$$tf\text{-}idf(|D|, w) = tf(d, w) \cdot idf(|D|, w) = 0.2 \cdot 0.3 = 0.06$$

Table 3.2 shows tf-idf values for some arbitrary words in the corpus D.

Table 1.3 Tf-idf Example

This example would result in the matrix $X \in \mathbb{R}^{4\times10}$, where entries cor- responding to the words "beautiful", "shining", "sun" and "weather" would bethe same as in the table above. Entries referring to the other words would be calculated in the same manner.

2.1 General Concepts

Sentence	*beautiful*	*shining*	*sun*	. . .	*weather*
The sun is shining brightly.	*0*	*0.12*	*0.06*	. . .	*0*
The weather is nice and beautiful.	*0.05*	*0*	*0*	. . .	*0.05*
It is a nice weather.	*0*	*0*	*0*	. . .	*0.06*
The sun is beautiful.	*0.075*	*0*	*0.075*	. . .	*0*

Machine learning is a field of study that by using algorithms and mathematical optimization imitates the way that humans learn. By analyzing data, computers can make predictions, decisions, or perform grouping of data samples also known as clustering. There are two main variants of machine learning, one of which is supervised learning where the data is labeled. The label is an additional column in a dataset that tells the algorithm what type of observation is at hand. In the dataset used in this thesis, this column represents the innovation status of a company In order fora machine learning algorithm to learn, the dataset is separated into training and test sets. Using the training set, mathematical optimization is performedto make predictions about the data (Mohri et al., 2012). The performanceis subsequently evaluated by using the previously trained function to predict the output associated with new inputs from the test set. On the other hand, unsupervised learning does not have a label and usually clustering is the main focus of that category of machine learning. Cluster analysis attempts to put observations into clusters that contain similar data points according to pre- determined criteria. In this thesis, the algorithms used employ supervised learning techniques as labeled datasets are provided.

In addition to regular supervised

classification algorithms, ensemble learning will be employed. Homogeneous ensemble algorithms combine predictionsseveral weak learners into one single, improved prediction. This final pre- diction is made by a meta classifier, also called strong learner. Meta classifier is a separate model that as an input data uses outputs of models used in the ensemble. Weak learners are small, single-type algorithms that perform relatively poorly and their accuracy is usually close to a random chance (Elishetal ., 2013). A strong learner iterates over weak learners in an attempt to weigh and aggregate them resulting in a better classification. Common weak learners that are used in many algorithms are decision trees. A decision treeis a simple machine learning model that makes predictions based on the mapof the possible outcomes of the sequence of choices. Two instances of decisiontrees are presented below:

A simple example of homogeneous classifier is decision tree classifier presented in

Figure 2.1 It uses simple maority voting (hard voting) as a meta classifiers and five trees as weak learners

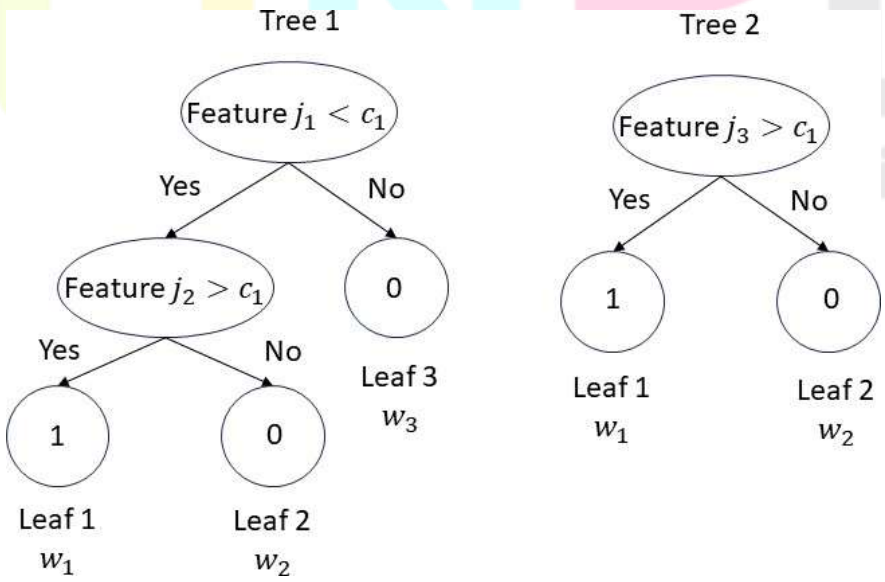


Figure 2.1: Example of decision trees

2.2 Logistic Regression

Logistic regression is a statistical model that predicts the binary outcome based on prior observations in the dataset. It allows to obtain probabilitiesof samples belonging to a particular category. Consider $p = W$ explanatory variables (features) and regression coefficients (predictors) β modelled as vectors, where

$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots x_{ip}) \qquad \forall i = 1, \ldots, |D|$$
$$\boldsymbol{\beta} = \{\beta_1, \beta_2, \ldots \beta_p\}$$

Logistic regression uses the logit function that ensures that the output isa real number between 0 and 1. Understanding the mathematics behind the logistic function is essential for understanding the logistic regression (Jameset al., 2013). The logistic function $\sigma : \mathrm{R} \rightarrow [0, 1]$ is defined

Logistic regression aims to predict the probability of $y_i = 1$ given $\boldsymbol{x}_i$. Themodel is formulated as follows:

$$P(y_i = 1|\boldsymbol{x}_i) = \pi(\boldsymbol{x}_i) = \frac{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}}{1 + e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} \quad \forall i = 1, \ldots, |D|$$

Furthermore, the following transformation can be applied to obtain thelinear predictors ($\beta_1, \ldots, \beta_n$).

$$\log \frac{\pi(\boldsymbol{x}_i)}{1 - \pi(\boldsymbol{x}_i)}$$

$$\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

$$\forall i = 1, \ldots, |D|$$

The coefficients β are unknown. They need to be estimated using the training data. For Linear Regression, the coefficients are estimated by minimizing the sum of squared deviations of the predicted values from the observed values.This deviations are also called residuals. In logistic regression a transformationexplained in Equation as used that apart from obtaining linear predictors, transformed the y-axis being probability $\pi(\boldsymbol{x}_i)$ to + or Least-squares method is unable to find the best fitting line as residuals would be close to infinity as it can be seen

Calculating squared distances from these

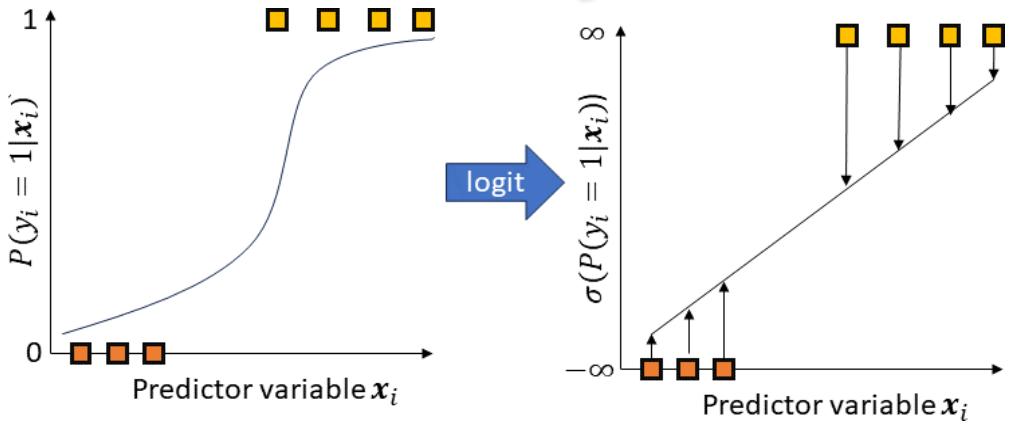


Figure 2.2 Logit transformation in Logistic Regression

points (residuals) to the logistic function line is impossible because ofthe applied transformation. This means that a maximum likelihood estimation is needed to find the values of β (Czepiel, 2002). By projecting data points on

the likelihood function presented below, log odds are calculated that can rep- resent how accurately the maximum likelihood function represents the actual probabilities. By using

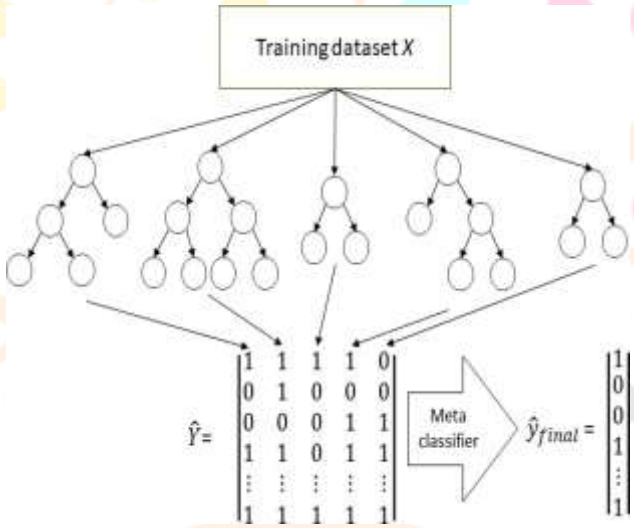
the following function with the best possible β values, the algorithm provides accurate log odds for the data.

$L(\beta) = \prod_{i=0} \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i}$ is estimated by maximizing the function above. One technique that can be erformed here is equating every partial derivative of the above equation with respect to each of the β to zero. It will result in a set of equations that will hold at

the maximal value of the log-

Figure 2.3 Homogeneous ensemble classifier example - hard voting

likelihood. $\frac{\partial L(\beta_j)}{\partial \beta_j} = \sum_{i=1}^{|D|} y_i x_{ij} - \pi(x_i) x_{ij} \quad j = 1, \ldots, |W|$



her coefficients $\hat{\beta}$ can be interpreted as those influencing innovation more. The presence of words with negative coefficients means that a company is lessinnovative.

In addition, heterogeneous ensemble algorithms will be considered. They blend predictions of various different models, possibly including homogeneousensemble models into a large combined prediction by using a meta classifier. Hence, both homogeneous and heterogeneous models use a meta classifier to combine predictions of numerous classifiers. However, homogeneous ensemble methods have the same feature selection with distinct training data and al- locate the dataset over various nodes. Heterogeneous ensemble methods use disparate feature selection methods with identical training data. Feature selection is a process of selecting a subset of relevant features that include variables and predictors to improve the performance of an algorithm. For example, homogeneous ensemble method random forest uses bagging to select a subset of data for every weak learner (tree) within the model. Heterogeneous models combine predictions of various different models that them selves use various feature selection techniques. Naturally, this means that feature selection is not the same for every model included in the ensemble. Random Forest employs bagging as a feature selection technique but Logistic Regression contains Lasso penalty. The models included in the ensemble will contain the most accurate models found by Daas and vander Doef (2020), namely Logistic Regression, Support Vector Machines, and Neural Networks. Random Forestswere found to be a decent method by Peerboom (2023) when using sentence encoding instead of *tf -idf* . Despite lower accuracy when using *tf -idf* , it can still hold valuable information. Additionally, two new models will be included,namely Adaboost and XGBoost. Boosting methods provide an interesting alternative to Random Forests as they attempt to weigh the trees to reduce thenumber of misclassified observations of the classifier.

REFERENCES

1.Abbasiharofteh, M., Kruger, M., Kinne , J., Lenz, D., and Resch, B. (2023). The digital layer: alternative data for regional and innovation studies.*Spatial Economic Analysis*, 0(0):1–23

2. Aggarwal, C. C. (2014).Data Classification Algorithms and Applications. Chapman Hall/CRC, 1st edition.

3. Altaee , H. (2023). Webpage classification using ensemble machine learning.Iraqi Journal of Intelligent Computing and Informatics (IJICI), 2.

4. Ashouri, S., Suominen, A., Hajikhani, A., Pukelis, L., Schubert, T., Türkeli, S., Van Beers, C., and Cunningham, S. (2022). Indicators on firm level innovation activities from web scraped data. Data in Brief, 42.

5. Beel, J., Gipp, B., Langer, S., and Breitinger, C. (2016). Research paperrecommender systems: a literature survey. International Journal on Digital Libraries, 17(4):305–338.

6. Breiman, L. (2001). Random forests. Machine Learning, 45(1):5–32. Cameron, G. (1996). Innovation and Economic Growth. discussion pers,Centre for Economic Performance, LSE.

7. Chai, C. P. (2023). Comparison of text preprocessing methods. *Natural Lan- guage Engineering*, 29(3):509–553.

8. Chen, T. and Guestrin, C. (2016). Xgboost. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Min-ing. ACM.

9. Commission, E. and Eurostat (2008). NACE Rev. 2 – Statistical classificationof economic activities in the European Community. Publications Office.

10. Cutler, A., Cutler, D., and Stevens, J. (2011). Random Forests, volume 45, pages 157–176.

11. Czepiel, S. A. (2002). Maximum likelihood estimation of logistic regression models Theory and implementation.

12 .Daas, P. and Jansen, J. (2020). Model degradation in web derived text-based models.

International Research Journal
IJNRD
Research Through Innovation