Analysis of the Quality of the Painting Process Using Preprocessing Techniques of Text Mining

Veronika Simoncicova[1](✉), Pavol Tanuska[1],
Hans-Christian Heidecke[1], and Stefan Rydzi[2]

[1] Faculty of Materials Science and Technology in Trnava,
Slovak University of Technology, Bratislava, Slovakia
{veronika.simoncicova,pavol.tanuska}@stuba.sk,
christian.heidecke@gmx.de

[2] PredictiveDataScience, s. r. o., Klzava 31, Bratislava, Slovakia
stefan.rydzi@predictivedatascience.sk

Abstract. Text mining is a relatively new area of computer science, and its use has grown immensely lately. The aim is to join two dataset from different data sources and to acquire information about percentage defects from the painting process, which are transmitted from the manufacturing to the end customers. The data sets are totally different and for their joining using text attributes, preprocessing are needed.

Keywords: Text mining · Data set · Data · Defect

1 Introduction

Real-word data are often incomplete, noisy, uncertain, and unreliable. Information redundancy may exist among the multiple pieces of data that are interconnected in a large network. Information redundancy can be explored in such networks to perform quality data cleaning, data integration, information validation, and trustability analysis by network analysis [1].

There are many other kinds of semi-structured or unstructured data, such as spatiotemporal, multimedia, and hypertext data, which have interesting applications. Such data carry various kinds of semantics, are either stored in or dynamically streamed through a system and call for specialized data mining methodologies. Thus, mining multiple kinds of data, including spatial data, spatiotemporal data, cyber-physical system data, multimedia data, text data, web data, and data streams, are increasingly important tasks in data mining [1].

Text mining has recently come to the forefront, especially of researching text information from social networks, various text documents, but the spectrum of use is very widely.

Feldman [2] wrote, that text mining is a new and exciting area a of computer science research that tries to solve the crisis of information overload by combining techniques from data mining, machine learning, natural language processing,

R. Silhavy (Ed.): CSOC 2018, AISC 764, pp. 30–38, 2019.
https://doi.org/10.1007/978-3-319-91189-2_4

information retrieval, and knowledge management. Similarly, link detection – a rapidly evolving approach to the analysis of text that shares and builds on many of the key elements of text mining – also provides new tools for people to better leverage their burgeoning textual data resources. The main tasks of link detection are to extract, discover, and link together sparse evidence form vast amounts of data sources, to represent and evaluate the significance of the related evidence, and to learn patterns to guide to extraction, discovery, and linkage of entities.

Data preprocessing is used on the different kind of data for example data from social network [3], data obtained from a laser confocal microscope [4] or surveillance of healthcare data [5] and many various another data. We decided to apply it, especially several techniques for data pre-processing, on painting process data and data from authorised service with the aim of analysing the quality of the painting process in one of the automotive company.

2 Preprocesing Techniques of Text Mining

As the text data often contains different formats like number formats, date formats and the most common words unlikely to help Text mining such as prepositions, articles, and pro-nouns. These can be eliminated by the pre-processing techniques. These techniques eliminate noise from text data, later identify the root word for actual words and reduce the size of the text. We know several different pre-processing methods, for example, tokenization, stop word removal and stemming for the text. Methods of stemming and stop wording were used by the quality analysis, therefore, they will be described in more detail.

All changes will contribute to better, faster and more efficient analysis of data. Preprocessing the data is very important part of data pre-processing, as the quality of acquired results depends on the quality of the used data [6, 7].

2.1 Stemming

Stemming, this method is used to identify the root/stem of a word [8]. For example, the words argue, argued, argues, arguing, all can be stemmed to the word “argu” (Fig. 1).

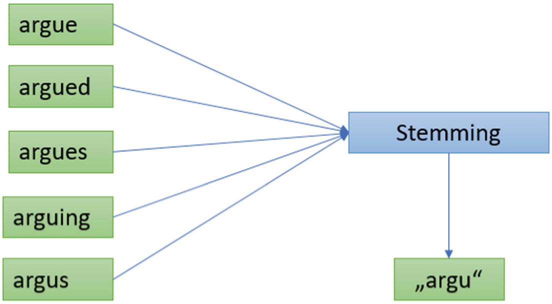

Fig. 1. Stemming process

There are mainly two defects in stemming [9]:

- over stemming,
- under stemming.

Over-stemming is when two words with different stems are stemmed to the same root. This is also known as a false positive. Under-stemming is when two words that should be stemmed to the same root are not. This is also known as a false negative.

2.2 Stop Word Removal

Many words in documents recur very frequently but are essentially meaningless as they are used to join words together in a sentence. It is commonly understood that stop words do not contribute to the context or content of textual documents. Due to their high frequency of occurrence, their presence in text mining presents an obstacle in understanding the content of the documents. Stop words are very frequently used common words like ‘and’, ‘are’, ‘this’ etc. They are not useful in classification of documents. So, they must be removed. However, the development of such stop words list is difficult and inconsistent between textual sources. This process also reduces the text data and improves the system performance. Every text document deals with these words which are not necessary for text mining applications [9].

3 Analysing the Dataset from the Manufacturing

In this section we will summarize the practices for collecting, integrating, interpreting, preprocessing and presenting data. The aim is to join two data sets from the different data sources and to acquire information concerning percentage of defects transferred from the manufacturing to the end customers. First dataset represents painting defects in manufacturing and the second dataset stands for records from the authorised services, where the claimed painting defects from the produced and sold products are stored.

These datasets contain description of defects, description of their location, but their quantity is very different, dataset from the service contains hundreds of records and dataset from the painting shop contains millions records about painting defects from the manufacturing.

Currently, these datasets cannot be compared by the Department of Quality. We work with useful software for analysis data, Rapidminer, to solve this issue, which will be described in the next chapter. Text information is the key attribute are, since we used several text mining techniques for the text preprocessing.

3.1 Rapidminer

Rapidminer is currently one of the most used an open source predictive analytics platform for data analysis. It is accessible as a stand-alone application for information investigation and as a data mining engine for the integration into own products. Rapidminer provides an integrated environment for data mining and machine learning procedures, including [10]:

- extracting the data from different source systems; transforming the data and loading into a data warehouse (DW) or data repository other applications,
- data pre-processing and visualization,
- predictive analytics and statistical modelling, evaluation, and deployment.

What makes it even more powerful is that it provides learning schemes, models and algorithms from WEKA and R scripts [10].

Rapidminer offers a large amount of different operators, which can be easily extended using existing extensions. There are packages for text processing, web mining, weka extensions, R scripting, series extension, python scripting, anomaly detection and more [11].

The Rapidminer text extension adds all operators necessary for statistical text analysis. Text from different data sources can be loaded and can be transformed by different filtering techniques, to analyses text data. The Rapidminer text extensions support several text formats including plain text, HTML, or PDF. It also provides standard filters for tokenization, stemming, stop word filtering, or n-gram generation. The Text Processing package, can be installed and updated through the Market Rapidminer menu item under the Help menu. The Text Mining extension uses a special class for handling documents, called the document class. This class stores the whole document in combination with additional meta information [11].

In an industrial company, there is large amount of data sources. In order to carry out our task, we will use two data sources, an internal data source for production data and an external source for field data respectively data from the market. All numeric values were changed in accordance with the company's security policy.

3.2 Dataset from Autorised Services

The system with field data is used to analyse the quality of the produced products. In that system, service data is stored, where the various attributes detail the reported claimed defects that have been reported to customers at authorised service centers. The unique object or product number (UVN), name of defect, the wrong part of the product, the exact location defects on the product, the type of product, production date, type of defect, the mileage and other attributes that accurately describe defect on the claimed

Tab. 1. The example set of service dataset

UVN	Wrong part	Location	Type	Production date	KM	Type of defect	Name of defect
11111	Front bonnet	Right	Sedan	12.10.2016	3215	Paint	Paint scratches
22222	Door	Left	SUV	16.06.2015	20085	Paint	Paint scratches
33333	Back bonnet	Front left	Kombi	07.11.2016	15	Paint	Skimmed place
44444	Door	Front right	SUV	08.07.2016	715	Paint	Spotting
55555	Door	Back left	Kombi	11.05.2015	0	Paint	Paint scratches
66666	Mirror cover	Right	SUV	06.03.2015	7	Paint	Paint scratches

products, are the most important attributes. Table 1 represents a sample of data from the services. For us, text is the key attribute. However, it requires to be preprocessed, as sometimes the same defect differs in name

Data analysis is currently performed manually and only on current data in a time interval of up to one month. The data, we obtained, concern products belonging to the year 2015 and 2016. There are exactly 274 records from this period. The number of products with defects is 174, containing different types of products. We analysed only painting defects from manufacturing.

3.3 Dataset from the Painting Shop

The production system database is used to write painting defects in production, but only a part of the data can be obtained from this database. A data sample of the painting defects from production is shown in Table 2, where we can see the attributes such as unique object number, defect description, equipment, object type, occurrence time of an defect, and number of repetitions, indicating how many times the object/product has returned to the repair.

Tab. 2. The example set of painting defects

UVIM	Error description	Equipment	Status	TN	Type	Ocurrence time	Repetiton
11111	Front bonnet right scratcher	HJ444	LQ00	92	Sedan	01.12.2016	0
10000	Door left front painting scratcher	HJ333	LQ00	92	Sedan	01.12.2016	1
23555	Back bonnet spotting	HJ444	LQ00	88	SUV	01.12.2016	1
44444	Door right front skimmmed places	HJ222	LQ00	88	SUV	01.12.2016	0
66666	Front bonnet right skimmed places	HJ333	LQ00	88	SUV	01.12.2016	0
88888	Door back left painting scratcher	HJ111	LQ00	72	Kombi	01.12.2016	1

The dataset contains painting defects directly from the paint shop as well as painting defects recorded in the assembly line. This dataset, we added based on painting experts' recommendations. The total number of records is 4.5 million for the period from January 2015 to March 2017. The number of objects with defects is 440 thousand.

3.4 The Process of Analysis Datasets

The first step was to select the test example set from the dataset from painting shop using “Sample“ operator, since the amount of data has a great demand on computer memory. The two datasets were joined using a unique identifier, unique vehicle number, where, there is the example set of each kind of data and joining these two datasets new dataset was created, which is located on the right of the figure. We identified objects occurring in both datasets. The main issue stemmed from the incorrect description of a wrong part of the object and from non-exact location of the defects on object/product, not matching the defect description.

Dataset containing only those examples where the key attributes of both input example sets match is the result of the first phase. The total number of matching records is 1225. "Join" operator connects records to each other, i.e. if there are 2 records of the same in the service dataset and 4 records stemming from painting, 2 * 4 will result in 8 entries.

In the second phase, the problem of different defect names had to be solved and at the same time, only totally or partially identical ones were to be selected. Defect descriptions were different in each dataset, and there were various problems in comparing these attributes, for example, a different format of defect and placement names, different naming, diacritics, and others. These defects were removed using Text mining methods for text editing; stemming and stop word techniques were used. We worked with Rapidminer software, "Stem (Snowball)" operator was utilised. This operator stems words by applying stemming algorithms written for the Snowball language. Various stemming algorithms for different languages can be chosen. "Filter Stopword" was used to remove stopwords from a document.

After this adjustment, the second phase, shown in Fig. 2, was carried out. In the Figure, individual data sets already been linked based on the primary identifier are depicted. In this file, we modified text attributes using pre-processing techniques of Text mining, and by comparing text we determined the complete or partially consistent records.

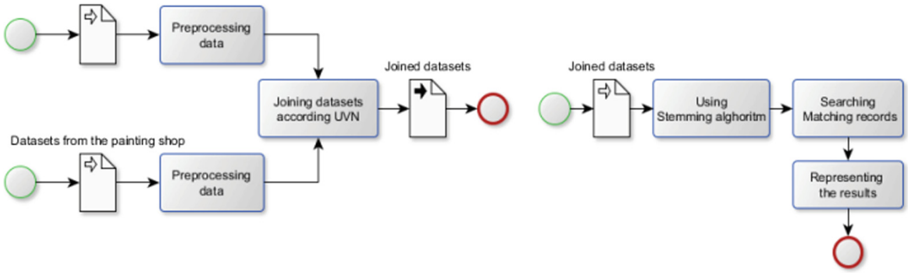

Fig. 2. The process of the modification of text attributes

As a result of these adjustments, we identified exactly 40 records from a total of 274 entries; all matches found being product A, which could be explained by having only 2% of product B defects in the painting data, and the rest 98% concerning the product A. The total number of products matched was 24, i.e., 8.8% of the total number of products recorded in the service (174 objects/products). These were the first relevant results. Consequently, they were checked by the quality department, and following the consultations and their recommendations, we adjusted the process and received a percentage of 10% of the total number of tests recorded in the service.

Our accuracy of identifying the same mistakes was about 80%, which a very impressive result, as the fact that not all matching records had to be real and the same has to be taken into account. It is necessary to amend that sometimes an defect occurs just as a coincidence and this deficiency cannot be effectively eliminated. The results were checked by the quality department using a program to analysis all complaints, including photos.

The entire process of editing and identifying the particular disorders transmitted to a particular consumer is shown in Fig. 3.

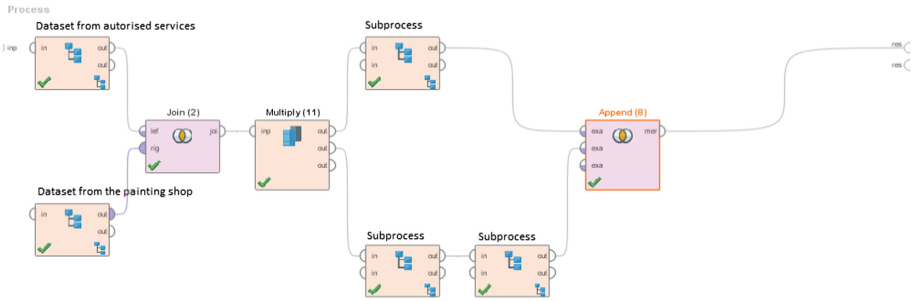

Fig. 3. The process of analysis of quality

The process consists of retrieving data that had to be partially edited, i.e., it was necessary to modify data types, create a new attribute, and edit the names of some attributes. The next step was to associate data sets with the join operator, and the last step was to define totally identical records using modified text attributes and their cross-referencing.

4 Conclusion

The main objective was to make efficient and to improve the analysis of the quality of product painting in the particular industrial company. The Department of Quality currently works only with service data that has been evaluated from a variety of perspectives, and when an increased incidence of a particular defect is recorded, the painting process is to be analysed and corrective actions can be taken. The dataset obtained from services is not possible to be compared with such a large amount of data from the painting shop. Text analysis and the proposed process helped expanding the view of the quality of the produced products and streamlined and accelerated work; the entire process of retrieving new data and comparing it takes about few minutes.

The all stages are depicted on the Fig. 4, where is:

- Identification the aim - clear and exactly defined aim is important for process of data analysis,
- Obtaining data sources - this process is very complicated in a large company and it is therefore necessary to identify and collect all necessary data for analysis,

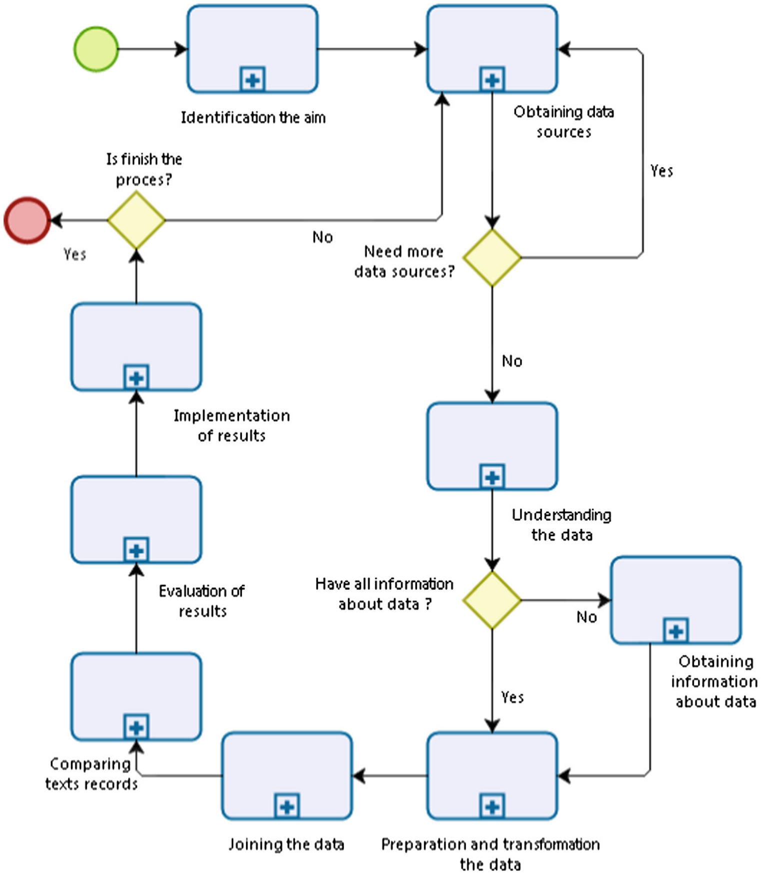

Fig. 4. The process of analysis of quality

- Understanding the data - find detailed information about the data
- Preparation and transformation the data the longest and most complicated stage of the whole process, this stage affects the final results,
- Joining the data - the basis is to find the primary key to be linked,
- Comparing text records - the main purpose is to find exactly identical or partially identical text records,
- Evaluation of results - a summary of the results achieved,
- Implementation of results.

These results will be further used to extend the process on the longer time frame of painting defects. We are currently working on data processing from 2011 to 2015, when changes were made to the names of defects and locations and the structure needed to be unified. We also plan to use some datamining methods and other text mining tools. Later, we plan to extend the dataset to other defects, stemming for example from engine, electrical issues and others.

Acknowledgment. This publication is the result of implementation of the project: “Increase of Power Safety of the Slovak Republic” (ITMS: 26220220077) supported by the Research & Development Operational Programme funded by the ERDF and project VEGA 1/0673/15: “Knowledge discovery for hierarchical control of technological and production processes” supported by the VEGA.

References

1. Han, J., Kamber, M., Pei, J.: Data Mining Concepts and Techniques, 14 February 2018. https://books.google.de/books?hl=sk&lr=&id=pQws07tdpjoC&oi=fnd&pg=PP1&dq=data+mining+text+mining&ots=tzEy0-pzX2&sig=3y8SbPuEoEeYkbE8A69jA2st890#v=onepage&q&f=false
2. Feldman, R., Sanger, J.: The text mining handbook, 14 February 2018. https://books.google.de/books?hl=sk&lr=&id=U3EA_zX3ZwEC&oi=fnd&pg=PR1&dq=feldman+text+mining&ots=2NxKMiDwOG&sig=hDTiHAMhaeJ83NzmtS8CME4PmZA#v=onepage&q=feldman%20text%20mining&f=false
3. Domingos, P.: Mining Social Networks for Viral Marketing, 14 February 2018. http://ncwebcenter.com/domingos05.pdf
4. Bezak, T., Elias, M., Spendla, L., Kebisek, M.: Complex roughness determination process of surfaces obtained by laser confocal microscope, 14 February 2018. http://sci-hub.hk/, http://ieeexplore.ieee.org/abstract/document/7555111/
5. Obenshain, M.K.: Application of Data Mining Techniques to Healthcare Data, 14 February 2018. http://sci-hub.hk/, https://www.cambridge.org/core/journals/infection-control-and-hospital-epidemiology/article/application-of-data-mining-techniques-to-healthcare-data/7EE5E7B1FA8B1C535FBC7A3881EC42
6. Simoncicova, V., Hrcka, L., Tadanai, O., Tanuska, P., Vazan, P.: Data Pre-processing from Production Processes for Analysis in Automotive Industry, 14 February 2018. http://archive.ceciis.foi.hr/app/public/conferences/1/ceciis2016/papers/DKB-3.pdf
7. Ramasubramanian, C., Ramya, R.: Effective preprocessing activities in text mining using improved porter’s stemming algorithm. Int. J. Adv. Res. Comput. Commun. Eng. **2**(12), December 2013. ISSN (Online): 2278-1021
8. Gurusamy, V.: Preprocessing Techniques for Text Mining, 20 July 2017. https://www.researchgate.net/publication/273127322_Preprocessing_Techniques_for_Text_Mining
9. Gupta, G., Malhotra, S.: Text documents tokenization for word frequency count using rapid miner (taking resume as an example). Int. J. Comput. Appl. (0975-8887). International Conference on Advancement in Engineering and Technology (ICAET 2015) (2015)
10. Akthar, F., Hahne, C.: RapidMiner 5 Operator Reference, 20 July 2017. https://rapidminer.com/wp-content/uploads/2013/10/RapidMiner_OperatorReference_en.pdf
11. RapidMiner text mining extension, 20 July 2017. http://www.predictiveanalyticstoday.com/rapidminer-text-mining-extension/