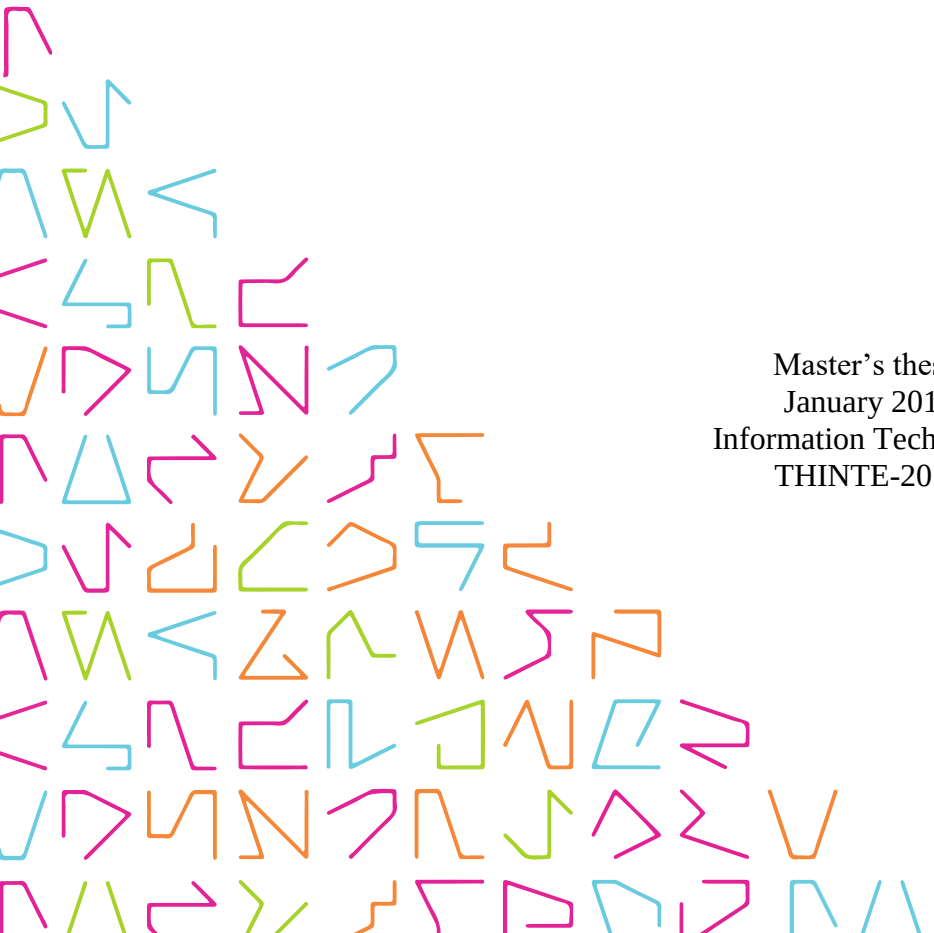
TAMPERE UNIVERSITY
OF APPLIED SCIENCES

Automated Analysis of Feedback in Various Formats of Data

Search for methods and tools to extract insights and information

Harri Huotari



Master's thesis
January 2017
Information Technology
THINTE-2016

ABSTRACT

Tampereen ammattikorkeakoulu
Tampere University of Applied Sciences
Information Technology, Master's thesis
THINTE-2016

HUOTARI, HARRI:
Automated Analysis of Feedback in Various Formats of Data
Search for methods and tools to extract insights and information

Master's thesis 48 pages, appendices 8 pages
January 2017

Jaxber is a cloud-based tool which collects feedback, sharable information and ideas in various formats in targeted surveys or campaigns. The feedback is collected by a mobile application. The content can be text, videos, audio files and images in multitude of languages. Jaxber is owned, developed and marketed by Nestronite Oy company.

The purpose of this thesis was to find tools which can process all the content produced by Jaxber in as automated manner as possible. The urgency to shorten the time spent going through the feedback increases when the amount of material gets overloaded. The processing was to provide summaries, which highlight the essential information of the feedback for the customer of the campaign.

Potential methods and tools for the task were studied from the market mostly by browsing the offering of different IT companies and open source providers. The most promising ones were experimented with two sets of real data collected by Jaxber. Free trials of SW products and development environments were available in most cases for assessing suitability.

No single tool was found for the task. The most promising one for natural language processing was Bitext product. For extracting information from video files, Google Speech API is recommended, partly due its support for 80 languages. The process involves transforming the content first to audio files, and then transcribing it from speech to text. Google's Vision API can be used for analyzing image content.

This is only the start. The trials prove that the technology is existing. Now it is time to develop necessary application for automating the processing of input files and utilizing cloud based services. Python as a programming language is recommended as it supports open source libraries for further natural language processing. For text analytics, the Bitext can be taken into use at any time, if the price is found as acceptable for the company.

Key words: data analytics, natural language processing, automated analysis, qualitative data

CONTENTS

ABBREVIATIONS AND TERMS

AGC	Automatic Gain Control, one of the audio processing methods
API	Application interface, a pre-defined way to access services provided by an application
Corpus	A set of words related to any specific area of interest, or can be also a dictionary of a certain language. In plural Corpora.
cURL	A command line tool for transferring data by using various protocols. Originally the name was “see URL”.
EMR	Elastic Map Reduce, marketing name of Amazon for their own variant of MapReduce technology
Feedback	This paper uses the word ‘feedback’ to describe any piece of information, which is captured with Jaxber. Also, insights, ideas and information sharing.
Flac	A high-quality audio format, free lossless audio coding
GB	Gigabyte, 1000 000 000 bytes of data (1024 x 1024 x 1024)
GCP	Google Cloud Platform
h2o	R project extension to big data resources through services provided by h2o.ai
HDFS	Apache Hadoop Distributed File System
HP	Hewlett Packard, a global IT company selling devices, software and services
IDE	Integrated development environment
Jaxber	Mobile application to capture insight and feedback
Linear16	Google’s abbreviation for an uncompressed headless raw audio format with signed 16-bit words, pulse code modulated (PCM)
LREC	Bi-yearly Conference on Language Resources and Evaluation
MB	Megabyte, 1000 000 bytes of data
Nestronite Oy	A company owning Jaxber and providing services with it
NLP	Natural Language Processing
NLTK	Natural Language Toolkit, open source software for language processing with Python
Nvivo	A software tool for qualitative analysis of textual data

Python	Programming language, compiled on run time, used e.g. for high-level scripting, open-source with a lot of libraries
R	R-Project for statistical computing, a language and community developing software tools
SAS	SAS Institute Inc., an US based IT company
Sodexo	A catering company for which Jaxber has been used to collect feedback, used as an example
SoX	A command line audio utility tool to convert various audio files to different formats, SoX = **So**und e**X**change
tm	text mining package of R-project

1 INTRODUCTION

Jaxber (Nestronite Oy, 2016) is a mobile cloud-based application to capture and share insights, feedback and knowledge. Throughout this thesis insights, feedback and sharable data are called as feedback for the sake of simplicity. The purpose is to study possibilities for automated analysis of unstructured content, which is provided by feedback campaigns carried out with the help of Jaxber application.

The data format can be text, audio, video or images. The thesis concentrates on primarily on textual content, but possibilities to extract information from video, audio and images are also considered. The theory behind intelligent analytics of textual data is introduced for some of its parts. Commercial tools are experimented on their suitability for analysis. Also, it will be studied what open-source methods and tools are available.

Most promising tools and methods are experimented with the real data produced by one of the earlier campaigns with Jaxber. Finally, recommendations will be given on the possible tools, and how the work could be continued within Nestronite Oy, which is the company owning the Jaxber tool and the concept behind it.

2 Jaxber and goals for the analysis

This chapter describe what Jaxber is and what kind of use cases it enables. The goals for processing the output data are described, too.

2.1 Jaxber and its usage

Jaxber itself is a free mobile application, developed for Android and iOS operating systems. For using it, an agreement needs to be done with Nestronite to initiate a campaign. The data collected with Jaxber is stored to cloud. Amazon and DigitalOcean have been used as providers of cloud services.

Most commonly known tools use forms with predetermined questions and alternatives for collecting feedback. This is not the case with Jaxber. The idea is to collect open unstructured data of any format. This puts pressure for analysing and realizing rapidly what was the real content of the feedback. At the same time, the requirement for the intelligence of the analysis is increasing. If the size of the campaign is considerably large in terms of responders, analysing task is very laborious manually. This is exactly where this kind of automated analysis is called for to deliver a report of essential findings in time to the customer who ordered the campaign. Jaxber supports several languages (Finnish, English, Polish and many more).

2.2 Use cases of Jaxber

The extend of possible use of Jaxber is large. E.g. it can be used internally within a company collecting insight and ideas from the employees or externally to collect feedback from the customers. It can be used by marketing, human resources, educators, communications, development teams and management. It can be used also for instant sharing of data within a project group and its stakeholders regardless of their location. The data can be configured as visible for the members of the campaign instantly. The fact that it is a mobile application also provides more unexperimented possibilities for the teams to communicate when they are on the move. E.g. it was used in an international business forum event to capture improvement ideas and to create a collective diary on the interesting pieces of information by several participants of a single company. There are number of ways to configure the campaign based on the needs of the customer.

2.3 Context and goals for the analysis

The desired functionality of the automated analysis tool is extensive. This thesis does not even try to support all of them. However, they are listed in this chapter to understand what would be a complete solution in longer term. The main purpose of the automation is to reduce the time and manual effort to go through the feedback data and grasp the essential content of the data.

2.3.1 Context of the automated analysis

The context is shown in figure 1 inside the red line.

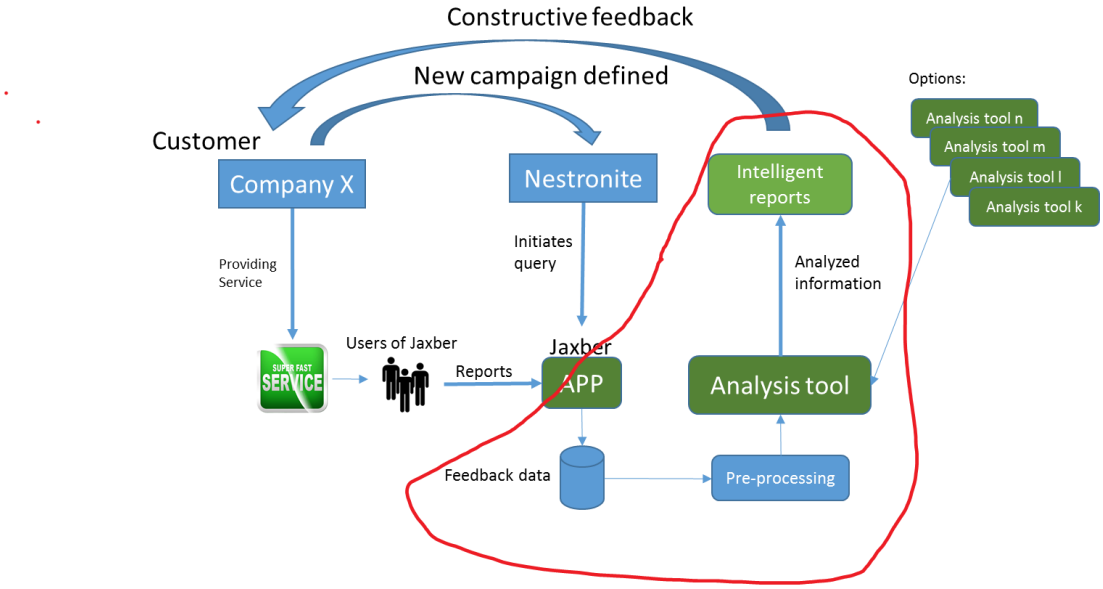


FIGURE 1. Context of the automated analysis

This is an example where Jaxber is utilized to provide feedback of a service. Company X orders the feedback campaign, which is initiated by Nestronite. Customers of Company X provide feedback by using Jaxber application. The analysis tool will provide summary of the feedback data to Company X. This is an ideal situation for a full-blown analysis tool.

2.3.2 Goals

These are the original goals, as listed in March 2016.

a) Similar feedback identified

- The analysis tool should reveal such topics which are being repeated in majority of the user feedback

b) Clarity of the feedback

- The outcome should be shown in a clear manner, at least textually and preferred also visually

c) Intelligent conclusions

- Optionally, the analysis tool should demonstrate trends or draw intelligent conclusions with the help of advanced algorithms on what the users are expecting from the service they are giving feedback

d) Several languages supported

- The analysis tool should be able to tackle the feedback in English and on top of it in as many European languages as possible (Spanish, German, Finnish, French, Italy, Swedish, Russian, Polish, …)

e) Ready by the end of 2016

- The tool need to be in use by the end of 2016, the scope need to be adjusted accordingly

f) Textual and audio formats

- Support for textual input is mandatory. Audio should be converted to text at least in case of English

g) Image and video formats optional

- Analysis of images and videos is an optional requirement. It can be regarded as a research topic for future development

h) Compatible with Jaxber

- The interface to Jaxber should be made in a way having reasonable effort on Jaxber side to make database accesses compatible

i) Open-source SW tool selected

- The development work will experiment open-source software tools and make a recommendation as an affordable (or free) tool for the automated analysis

j) Advanced SW tool selected

- The development work will experiment and provide recommendations or options as more advanced commercial software utilizing advanced algorithms

k) Easy to deploy, use and install

- The tool should be easy to use and easy to install to most common computer environments

l) Affordable cost

- The tool should be priced reasonable, and it should be able to run in most common computer environments

m) Large amount of data

- The tool should be able to tackle large amount of feedback data

n) Responsive usage

- The responsiveness of the tool should be reasonable from the customer point of view

o) Robustness to area of subject

- The tool should be able to deal with as many industrial areas as possible in consistent manner with the same performance level of intelligence

p) Great automated analysis provided

- Nestronite provides an automated analysis for its customers on the feedback data.

The list is so extensive that clearly these all cannot be fulfilled in the scope of a single thesis. Some of the goals are also contradicting compared to each other, indicated with red arrows. This is visualized in the figure 2.

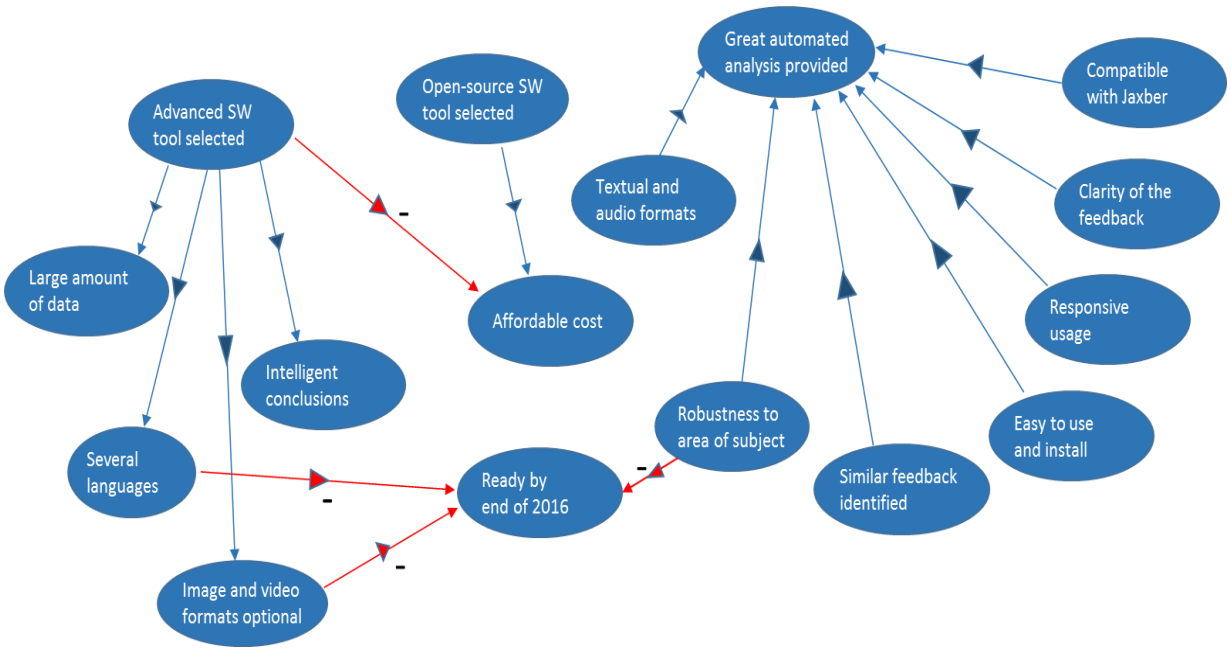


FIGURE 2. Goals visually with internal dependencies

3 METHODS FOR TEXTUAL ANALYSIS

What are the potential methods that the solution could use? What is the theoretical background? These are discussed in the following chapters. Traditional text analysis is described first. Then qualitative text analytics and machine learning are described as more automated methods.

3.1 Qualitative text analysis

There are a lot of open source and commercial tools for textual analysis and analytics. There are distinct differences in-between those ones classified as ‘analysis tools’ versus ‘analytics’.

The former one refers to old-school way of doing data analysis by the researches partly manually, and partly with the help of a tool. In such cases researches go first through the textual source material manually. Then they mark (term code is used also) certain parts of the text, categorizes them and combines them in many ways.

So, the analysis involves human interaction to tag terms, which are then subject for finding inter-relations and dependencies. The coding is very time-consuming and laborious task. In practice, it can be done only by the expert him/herself, having the specific area of interest and competence for it. So, the data preparation is not an automated task by no means, and nor are the tools for it in this kind of case.

The tools may support also video, image and audio content. Also, in case of video and image files the researcher needs to mark certain spots and code them to be used for further analysis and dependencies definition. Examples of these kind of free software tools are listed in reference (Predictiveanalytics, 2017). Examples of these kind of commercial tools are listed in reference (Predictiveanalytics C. , 2017). Nvivo is one of the latter ones, and it has been used for the analysis work of the data provided by Jaxber for some of the customers before this study was started.

3.2 Qualitative text analytics

Only seemingly minor difference in one word - analysis versus analytics. Here we get close to the header of the thesis. To simplify, automated analysis equals analytics. The

tools talking about analytics involve different machine learning techniques for understanding the content and making conclusions. Manual labelling or coding is not needed. The purpose is to extract high-value knowledge and information from natural language texts und ultimately present higher level ideas behind. Text analytics is used e.g. for the customer insight analysis, which is a typical case for the Jaxber. There are a lot of related techniques which are explained in the following sub-chapters. Some typical methods related are explained in Appendix 2. Text analytics is surely needed to be applied for the data provided by Jaxber. There are a lot of providers of commercial tools for this purpose [4]. The most known companies include IBM, Google, Amazon, Microsoft, Cisco, HP and SAS.

3.2.1 Text mining

Typically, the question is about searching information from the large source of data and documents for a specific purpose. It involves the ability to process unstructured text in such a way that it is possible to interpret the meaning of the source data, extract facts, relationships and assertions. As a contrast to search engines, which are associated to document retrieval, text mining is targeting for information extraction. Similar techniques are then applied as described above in chapter 3.2.

3.2.2 Natural language processing

Natural language processing (NLP) techniques are applied to extract feelings, attitudes or judgements from any source of textual data. The purpose is to identify and extract subjective information from language. Sentiment analysis has become one of the popular techniques to detect attitudes or judgements prompted by emotions.

When the source of data is coming from public sources like Twitter or other social media, the term opinion mining is used in this context. Nowadays Twitter is quite commonly used to reveal public opinions and sentiments. This is done typically with such tools that are having capability for big data, analytics and NLP. When all these applied for public data, status of any topic of interest, and even predictions to future can be achieved.

This was studied already 2010 in an article written into the Proceedings of LREC (Pak & Paroubek, 2010), where corpus of Twitter was analysed. The term opinion mining does not necessarily reflect the emotionally charged flavour, as with sentiment analysis. The

inter-relations and differentiation of the terms feelings, emotions, sentiments and opinions is discussed in the referred (Munezero;Montero;E.;& J., 2014) IEEE article. The proper understanding of these terms help automatic detection and processing.

3.3 Machine learning

Statistics and probabilities are forming the base for the machine learning. In the context of the term, it typically involves aspect of predicting the future or finding patterns from the existing data which leads to certain conclusions. Full data consist of data instances, which could be presented as feature vectors. Features are chosen with a specific task at hand. Based on the known data instances, one application is to make predictions on the future instances. Examples of the methods involved are classification, filtering, clustering and regression. This is not exact science, instead we are talking about probabilities (e.g. to decide whether a mail is a spam or not). The tools for data analytics benefit routines which are based on machine learning and the underlying statistics.

For building a clever process with the data analytics tool, it requires theoretical knowledge on machine learning and practical knowledge on the subject area. The pre-build statistical routines can be connected in a flow of processing the data. The cloud-based services can be part of the processing chain (obtained from e.g. Amazon, Azure or Google).

There are three main approaches (LoopAI-Labs, 2017) to machine learning which are supervised learning, semi-supervised learning and unsupervised learning – in the order of commonality. Supervised learning uses labelled training data for learning. The training data is prepared by a skilled human person, which is therefore costly. Semi-supervised learning uses only partly labelled input / output pairs. Unsupervised learning is based fully on intelligent algorithms to identify patterns in unlabelled data with very little or no human guidance.

Clearly, unsupervised learning is most difficult one and quite challenging, thus requiring the most sophisticated algorithms. This is exactly the case with Jaxber, as the textual data is natural language without pre-classification and training data for it.

3.3.1 Predictive analytics

Quoted partly from one of the IBM sources (IBM-SPSS, 2017): “Predictive Analytics helps connect data to effective action by drawing reliable conclusions”. Another from IBM: “Looking at the historical data, and by using modelling techniques to uncover patterns and relationships which are applied for new data for which the outcome is unknown”.

A bit tricky definition, but the basic idea is to predict from the past, and take then such actions which are favourable for the future business with high probability. E.g. a prediction can be made if there is a potential danger of a certain group of existing customers to resign. If so, then preventive actions can be taken through targeted marketing campaign to increase the attractiveness of a service or a product, to retain the existing customers.

That is an example where preventive actions are taken when a threat of losing business / customers is detected. This connection can be made automatic by linking real-time data to the process, and if triggered, inform marketing or sales organization on it. This kind of argumentation is very common for the tools listed in reference (Predictiveanalytics T. , 2017).

Typically, the tools assume that the input data is in a structured format or it first need to be converted into a desired format. Data sources may be also coming from cloud, on top of structured sources. There are certain data preparation techniques available for that purpose. Then data preparation could include filtering of unwanted data, and removal of insignificant stop words. The term ‘data cleansing’ is used often to describe the process of detecting, correcting and removing of corrupted, incomplete or irrelevant parts of the data (Wikipedia, 2017).

4 METHODS SUITABLE TO JAXBER

The previous chapter discussed about the methods more on the theoretical level. This chapter defines more specifically what kind of analysis is set as a target and what kind of methods should be applied.

4.1 The selected scope of analysis

Like earlier mentioned, the data formats of Jaxber are text, image, audio files and video files. The primary focus of this study is in the textual format, and such audio formats which can be transformed with sufficient reliability level to text. It would be a bonus, if video and image files could be analysed as well. Particularly it is studied, if the audio part of the video files could be extracted for further analysis. This means that also suitable tools for automatic speech recognition are studied. The purpose is to extract the essential information of the text:

- Repetitive words or themes, i.e. quantitative analysis
- Feelings, positive or negative
- Sentiment analysis (see chapter 3.2.2)
- Relations of words and themes
- Possibly embedded insights

If there is possibility to support several languages that will be considered as a major benefit, as the Jaxber is already now supporting 13 different languages, and 15 languages are under translation.

4.2 Needed methods for analysis

Considering the various formats of data that Jaxber is collecting, the analysing task is challenging. It clearly requires quite advanced tools and methods, as the content is unsupervised and mostly unstructured in nature. In other words, it is not known beforehand what we are looking for. The selected tool (or tools) are utilizing following methods and techniques:

- Statistical machine learning, modelling and computing
- Natural language processing
- Sentiment analysis
- Positive / negative detection

- Qualitative textual analytics
- Data preparation techniques

The need to transform video contents to text is also an important element while selecting the method and the tool as a lot of the input material collected with Jaxber are videos.

5 TOOL CANDIDATES

There are a big number of potential tools and methods. One of the sources for alternatives is predictive analytics web pages (Predictiveanalytics, 2017), which has been used in this study.

The alternatives are categorized here as follows:

1. Solutions with capability for big data analytics
2. Open source software platforms
3. Solutions of big IT companies for text mining and analytics
4. Free software for qualitative data analytics
5. Smaller companies with 'natural language first' approach

In practice the sources mentioned above were overlapping to some extend both in terms of the sources (companies and their products) and semantically (contents and purpose of use). Totally almost 200 service provides and tools were reviewed, some of them just briefly and some of them in more detail. The most promising ones for the task have been studied in more detail, as will be described in the following chapters.

Many tools utilize the open source software platforms as building blocks of their system. Many of the solutions are ready for the 'big data', which enable processing of huge amounts of data in parallel by the computer resources located in cloud, and are using also open source software architectures.

5.1 Solutions with capability for big data analytics

So, what is big data? The following list state 4 fundamental characteristics of big data:

- Volume: big amounts of data, e.g. order of terabytes of even 100-1000 terabytes
- Velocity: data in motion. E.g. streaming data. Anything that requires a rapid response or intelligent conclusions on-the-fly regardless of huge data amount
- Variety: data in many forms / formats / structures / unstructured data
- Veracity: data in doubt, uncertain data

Almost all the big data solutions provided by the number of companies are based on the technology originally released as a white paper by Google describing distributed file system and so called Map Reduce methodology.

The described method was developed as an open-source implementation in Yahoo's Apache project. The SW products hereby born were called Apache Hadoop, Hadoop's Distributed File System (HDFS) and Hadoop MapReduce.

Later it has evolved, and complemented by Apache Pig (the 1st high-level non-SQL framework for Hadoop), Apache Hive (Hadoop's 1st SQL access framework), Apache Spark (evolution of MapReduce boosting performance /speed significantly) and many more. Apache HBase is an evolution of basic HDFS. It is fault tolerant for large quantities of data including compression & filters.

This is not a full list of SW technologies around Apache Hadoop, but it gives a rough understanding on the topic. All the evolution has now lasted totally 10 years, counting from the initiation of Apache Hadoop project. The following figure illustrates Hadoop's distributed file system (HDFS).

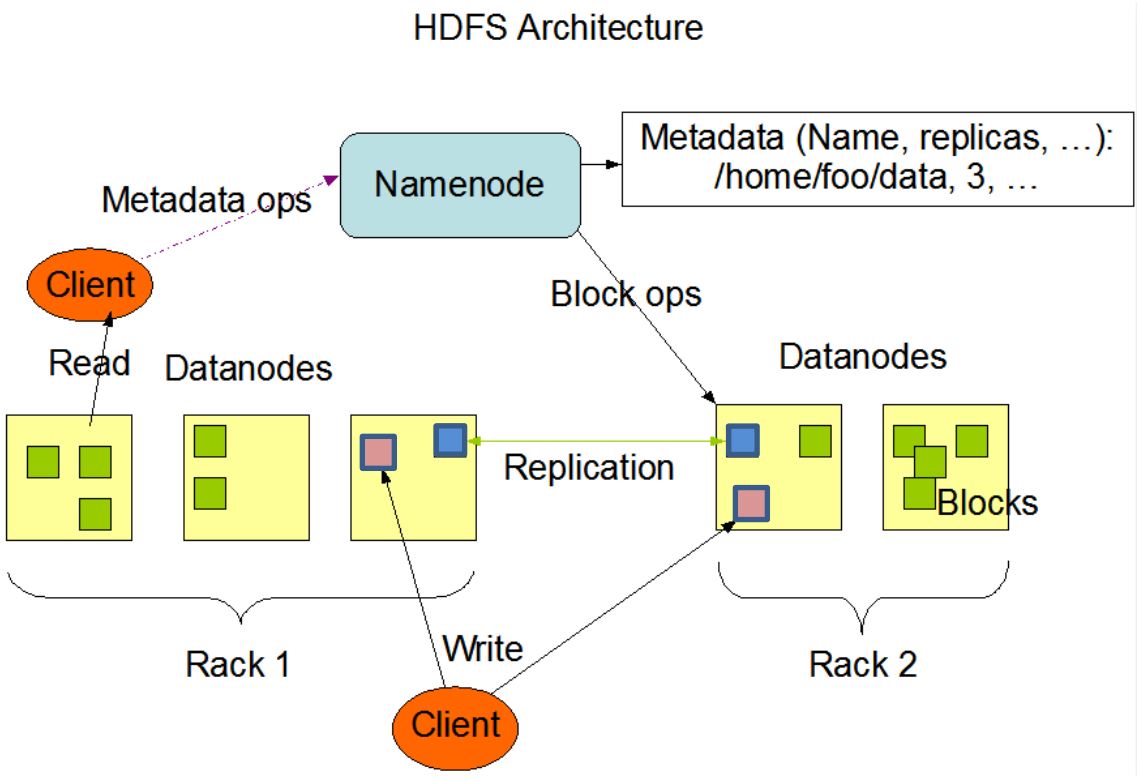


FIGURE 3. Hadoop's distributed file system architecture. Source (Hortonworks, 2017).

The data is typically replicated to 3 large blocks (typically 128 MB), and the blocks are located at least in two different physical racks to protect from external hazards. The racks may be in the same building or even in different cities, which may impact to the pricing of the service. Datanodes are consisting of processing elements with several processors and data storages. Namenode keeps track of the data blocks and their mapping to Datanodes. If a Datanode fails to respond or data is lost in one block this is detected by the Namenode.

Huge processing power is achieved through slicing the processing into several Datanodes. MapReduce arranges and slices the computing work as favourable for parallel processing. The number of Datanodes and data blocks can be increased for even more parallel processing on need basis. So, this is about a brute force of unlimited number of computer resources accompanied with local storage where latencies of accessing the data is reduced thanks to local copies of data.

Examples of file systems compatible with HDFS are Amazon S3 and Microsoft Azure, which are significant players in big data services and markets. Also, other big companies having big data analytics services are basing their system to open-source Apache Hadoop architecture and technologies (Oracle, Cloudera, Cisco, SAS, Hortonworks, Informatica, Alteryx and many more). Many of them have complemented it with their own innovation. E.g. Amazon is marketing EMR, i.e. Elastic MapReduce technology as part of the technology portfolio. Basically, it seems that Apache Hadoop has become an industry standard for big data implementations.

Apache Software Foundation (Apache, 2016) is an instance providing support for Apache Community of open-source projects aiming for the public good. It consists of users and developers with desire to create high quality software. There are over 100 projects listed under its umbrella, Hadoop being one of them.

5.2 Open source software platforms for data analytics

There are number of tools and software platforms as open source software implementations for data analytics. Many of them are utilized as a part of the commercial solutions.

5.2.1 Apache based platforms

Apache Mahout provides an environment for building machine learning algorithms. It provides routines for data mining clustering, classification, filtering and other machine learning algorithms and features for big data sets. The software provides three major features:

1. Programming environment and framework for building scalable algorithms
2. A wide selection of pre-made algorithms
3. Samsara, a vector mathematics based experimentation environment

Another one from Apache is OpenNLP. It is a machine learning based toolkit for processing natural language. It supports tasks that are usually required to build more advanced text processing services (Apache-OpenNLP, 2016).

Apache Lucene, although primarily targeted for indexing and searches, is also having advanced analysis and word tokenization capabilities, which could be utilized in this context.

Thinking about Jaxber, neither of these is a ready-made tool. They contain a set of library routines for text analytics. It is possible to develop a tailor-made tool, which would utilize those libraries, but it would take a lot of time and resources.

5.2.2 R Project for statistical computing

R is a language and environment for statistical computing and graphics (R-Project, 2016). It runs in most of the platforms, including Linux, Windows and Mac OS. It is a GNU project, and it is further developed from S language, which in turn was originally developed at Bell Laboratories. R provides an integrated suite of software for data manipulation and calculation and it includes:

- effective storage facility
- a suite of operators for calculations on arrays and matrices
- a large set of intermediate tools for data analysis
- graphical facilities for data analysis and display
- simple and effective programming language with loops, recursive functions and input and output facilities

- for computationally intensive tasks C, C++ and Fortran code can be linked and called at run time
- extensive and growing list of libraries, including data mining

R can be extended via packages. The list of domain areas of packages are listed here (R-Libraries, 2016). Natural Language Processing is one the domain areas, and it is further divided to tens of frameworks. The R Journal is a free publication having quite academic articles on the statistics and the progress in the area for the developers. There are also numerous books on it. Many commercial predictive statistical tools are referring to R as having support for extensions made with R. It can be used e.g. for text mining and text analytics. There are a lot of tutorials in YouTube how to use R for different tasks; e.g. Term Frequency and Word Clouds and another example of building a text mining, machine learning document classification system in R. One of the packages provide interface to OpenNLP tools.

In year 2014 (Revolutionanalytics, 2016), R was used by Facebook, Google, Twitter, New York Times, Microsoft, John Deere, Lloyds of London, RelaClimate.org, and many other software vendors including SAS, Oracle, IBM, Alteryx and SAP. This is not a full list, but gives an idea about its popularity.

RStudio is an integrated development environment (IDE) for R projects. Also, Microsoft is having their own 'release' of R, known as Microsoft R Open (v.3.2.5 in June 2016). There is both free open source and commercial version of Studio available.

Considering Jaxber, R has some benefits over Apache based analytics tools. RStudio is a decent development environment, and video training materials help to get started. The available library functions are also extensive. Many commercial tools have link to R functions, and so they are freely available. There is still long road ahead in using R for such sophisticated text analytics functions required by Jaxber. I spend one week of going through tutorials and exercises without getting yet to text analytics area. There are libraries and packages for text mining and analytics available. The most important one of them is tm (text mining), which is the core for other extensions of NLP. E.g. some basic routines like pre-processing, word frequencies, term correlations and clustering based on similarity of the terms are available (R-Studio, 2016). Packages can be easily installed to RStudio. Appendix 1 contains some further hints about how to experiment usage of R Studio.

5.2.3 Other open source software platforms

Natural Language Toolkit is a platform for building Python programs solving human language processing tasks (NLTK, 2016). It supports over 50 Corpora, such as WordNet and a suite for text processing libraries for classification, tokenization, parsing and semantic reasoning. There are other Python based libraries for text processing like TextBlob, Gensim, Pattern, Spay, Orange and Pineapple.

Stanford CoreNLP is also a suite for natural language analysis [18]. It is written in Java, and it supports Arabic, Chinese, English, French, German and Spanish. It is optimized to work best in English. It is licensed under GNU General Public License. There are wrappers for other programming languages like Python, Perl, nodeJS and R.

Freeling is another language analysis tool suite written in C++, and it is supporting variety of languages (English, Spanish, Portuguese, Italian, French, German, Russian, Catalan, Galician, Croatian and Slovene, among others).

5.2.4 Summary on open source software platforms

As a summary, using RStudio for Jaxber can be considered if there are time and resources to get deeper into the opportunities it would offer by own development efforts. Theoretical knowledge on statistics would be a benefit. For simple filtering and word clouds it is usable without any major effort.

R provide interface to OpenNLP, and R is utilized widely by the commercial tools. This tells about the popularity of it, and it is thus more future-proof solution, as it is evolving all the time. Other open source platforms are very much language-centric, but are lacking such popularity and advanced development environments as with R. R is having also package for h2o, which enables the processing to be distributed in parallel using big data resources (Freeling, 2016).

5.3 Solutions of big IT companies for text mining and analytics

A lot of big companies exist in the field. In many cases, it is not easy to get trial version of commercial software tools and solutions for text mining and analytics use. They'd rather to start business discussion directly aiming not only to provide products but services as well. The products and services are aimed to solve and improve sales and marketing type of issues - so very business oriented operation with high licensing cost, which would be quite costly to complement usage of Jaxber. They may not be eager to start discussions, if they see little potential for business with unknown small customers, and with that it is not easy to get deep enough understanding on the technology either.

Generally, a lot of marketing videos are available with testimonies of existing customers and the benefits they are gaining. There might be good videos amongst them from facts point of view, but it takes a lot of time to go through them. A trial version may or may not be easily available. The same go with the pricing – either it is available or not. In the latter case, you need to start business negotiations by contacting them to find out real costs.

Big companies like Amazon and Microsoft have productized their offering well. They might have smarter software modules for analytics available as a part of their offering, but they are not sold as a stand-alone product. This is valid also for smaller companies that it is difficult just to purchase the text analytics part of the offering. The above mentioned happened with SAS and HP.

So, in case of many companies and their solutions you should go with full speed and buy something to be able to do any trials to find out which more sophisticated features are supported. These companies provide their consultancy and experts to tailor their product to match better on the needs of the customer. This would be quite costly way to proceed in case of a small company like Nestronite. The challenge is to get real understanding on whether analytics of unstructured data is supported, and which data formats and languages are included.

5.3.1 IBM solutions

IBM Watson Analytics

This is either not for unstructured text analysis, but rather for business analysis of a structured data.

IBM SPSS Statistics

The abbreviation SPSS comes from the words Statistical Package for the Social Sciences, originally from the SPSS Inc., which was acquired by IBM in 2009. IBM SPSS Statistics is an integrated package of products supporting planning, data collection, reporting and deployment. It is aimed to increase revenue, conduct research and aid for decision-making. The premium version offers a variety of advanced statistical procedures and it is capable to model complex relationships of data sources. The input files need to be in structured format, and thus it is not suitable alternative for Jaxber.

IBM SPSS Modeler

IBM SPSS Modeler supports data modelling and entity analytics for to find hidden relationships. It is possible also to analyse social networks and their dynamics. It is capable for structuring the data, categorizing concepts, and showing their relationships with different visual reproductions. It applies linguistics NLP technologies. It is extensible to R and Python.

The cost of the solution is one of the criteria for the selection. I bought the tool with student licence of 123€. The textual parts of the Sodexo feedback were imported to the excel so that one column with description header had the actual text. That was indicated for the tool as an input. It is easy to generate different views with the tool. It can categorize text and detect dependencies on topics occurring frequently in the same context. It can be also configured to detect different type of text categories, like customer feedback as a one choice. Examples of views are in Appendix 4. One big drawback with IBM SPSS Modeler is the high price (Jan 2017). It costs over 11k€ per user per year with 12 months of technical support for Premium version, which has NLP features (IBM-SPSSModeler, 2017).

IBM Watson-enabled data platform

This was announced just recently in late October 2016 (IBM-Watson, 2017). It can deal with unstructured data, and it is built on Apache Spark with readiness to ingest more than 100GB data per second. The platform can be extended with SQL, Python, R, Java and Scala. It supports also RStudio. No pricing was visible yet, and it looks like this is overlapping with SPSS Modeler.

IBM solution for speech to text

IBM is also offering a service which provides API to transcript audio files into text and online conversion is also available (IBM-Cloud, 2017). It is part of the Watson Developer Cloud offering. The supported languages are US English, UK English, Japanese, Spanish, Brazilian Portuguese, Modern Standard Arabic or Mandarin. More languages are coming in near future. The application can be written in JavaScript or Python. First 1000 minutes are free and after that $0.02 per minute, on monthly basis. Considering needs of Jaxber at this stage it is cheap, because the free 16.7 hours period is sufficient on monthly basis.

IBM solution for image detection

IBM calls it Visual Recognition. It is part of the Watson cloud offering. It was experimented with a couple of images from Sodexo and Jaxber feedback. It detects objects quite well. Concerning text, it does not recognize the text in the display of a laptop, nor any text from any hand-written poster.

5.3.2 Google based solutions

Google Cloud Platform (GCP) is providing a wide variety of tools and services for computing, storing, networking, machine learning, managing and securing identities. Some of them are still in Beta or Alpha phase, but most of them are available as ready-made products. Out of those, machine learning services have the most potential to provide solutions to Jaxber requirements.

The core of machine learning platform is TensorFlow ™, which is an open source software library for machine intelligence. It is complemented with a bunch of software Engines and APIs. Particularly Vision API, Translation API, Natural Language API and Speech API (Google-Cloud, 2017) are of interest in this case. Those APIs could be used

to complement Jaxber tool selection for extracting information from e.g. the image and video content, and to respond to the requirement of multi-language support.

The Natural Language API provides a set of features for analysing unstructured text including sentiment. I did a short experiment with the Cloud Natural Language API (“Try the API”) with some extracts of text from Jaxber feedback. The output was a bit disappointing as not giving much information on the contents.

The video format has become popular in using Jaxber as a feedback tool. Google is having a Speech API, which is in beta phase currently, and everyone can try it for free [22]. It recognizes over 80 languages. As a part of this study an experiment was made about how the audio track from video files could be extracted and then feed through Speech API, which converts it into text. By taking this into use, a lot the time and manual effort could be saved compared to watching each video file individually and making notes of them. Extracting the audio track from video files (of mp4 format) was quite easy with VLC media player, which is open-source and very commonly used.

The Speech API supports Flac, Linear16, PCMU and AMR audio formats. In practice Flac and LINEAR16 are promoted as functioning the best. The latter one is basically signed uncompressed 16-bit wave format and no further audio processing in mono channel. Flac format is one of the modern lossless audio formats capable for high-quality audio reproduction, yet with decent coding efficiency of 30-50%. Flac is license-free. These two formats guarantee the best possible accuracy for speech detection. Apparently, Google’s algorithms are optimized for as natural speech as possible. Sampling rate of 16000Hz is recommended, and no further audio processing like AGC or noise reduction should be applied to the source file.

The pricing is based on the actual use of the services. For less than 60 minutes monthly usage it is free. For more than 60 minutes it is $0.006 per 15 seconds, which means roughly ~$1.5 for one hour of speech. The consumption based billing with that kind of price level sounds quite decent for Jaxber.

Google does not provide any integrated development environment (IDE) to utilize above mentioned APIs. It is possible to access the APIs through mobile devices (Android), web applications in PC or laptops, and other devices connected to the internet. The supported

languages include Python, JavaScript, Java and C#. So, for Jaxber a practical way forward would be to develop a web application, which would take all the feedback files as an input, process them and utilize Google Cloud APIs for getting the insight out of them. As a first step, which would benefit a lot Jaxber, would be to process video files into textual files, as described above.

Experimentation with Google Speech API

An experimentation was carried out to test the usage of Speech API. It could be done without an actual application. The needed steps for doing it are listed in Appendix 3.

The experimentation of Speech API was done first with a test file provided by Google. Once succeeded, it was due to be repeated with a real data extracted from the video files containing feedback on Jaxber. The tool has some limitations.

- For a synchronous request (immediate response), max length of the audio file is ~1 minute.
- For an asynchronous request, only Linear16 audio format is supported
- From time to time, the tool didn't respond anything, not even for the test file
- There are some limitations in audio formats. SoX command line tool was taken into use to convert audio files into the required formats. Linear16 was recognized being of the following format:
    - encoded as signed integer with 16-bit words
    - little-endian storing format for bytes in a word
    - mono channel
    - sampling rate of 16000Hz recommended
    - File needs to be of RAW type without any header information with extension of *.raw (wav -files are not supported)

If the implementation does not follow the format exactly as described above, then the algorithm does not work very well or not at all. After encountering and solving many smaller problems, I finally succeeded in getting something rational out of the videos recorded concerning the feedback of Jaxber itself.

The quality and correctness of the output is impacted of the quality of speech recording, how clearly and correctly it is pronounced, and the audio volume level, and what is the language dialect selection told to the translator. I needed to restrict the length less than

one minute, as only the synchronous method provided a response. The main phases of the transcription procedure were:

- transforming the mp4 video file into audio file. This was done for two video files; let us call them A-video and B-video. Both were first converted into flac format with VLC player. Flac format preserves the original content and nuances of the speech as natural for further processing. It is a lossless audio coding format (no information lost)
- transforming further the audio file of B-video into Linear16 -format. An example of 55 seconds was extracted from the flac file with SoX audio format conversion tool
- the audio file was copied into the Google Storage for each case. Before that you need to have Google Cloud account and Google Storage reserved for that purpose
- a json file was created, which specifies where is the data file (a location of the data bucked in the Google Cloud), and what is the format of coding. The details of json file are presented in Appendix 6.
- and finally sending the request to Speech API with 'curl' command as explained above with Google Cloud SDK Shell command line tool. For details, see Appendix 6.
- waiting for the response, which was received in json format to command line tool. For a 55 second audio file, it took roughly 70-80 seconds.

Speech API is an attractive option for Jaxber, and it is recommended to continue work to take it into use in full production. A better way to use it is to develop an application with Python, JavaScript, Java or C#. There are some examples of code written in those languages. The beta phase of Google Speech API has started around summer 2016, and hopefully it will reach soon the maturity level for official release.

Experiments with Google Vision API

Some images from Sodexo data containing images from the food plates were inserted into the Vision API. It detected quite correctly that there were food plates and with some accuracy what exactly (e.g. vegetables). When inserting hand-written textual posters related to the Jaxber group works, it had a lot of difficulties interpreting the text. There was one image taken directly from the display of a laptop (containing text), and that one it interpreted completely. The API can also analyse emotional facial attributes.

5.3.3 Alteryx analytics solutions

Alteryx provides tools for customer analytics, big data analytics and predictive analytics. Basic use case is to analyse sales data and perform analytics functions for the data. Alteryx's strength is consolidating data from external sources and cloud services like Amazon S3 and Microsoft Azure. Other partners were mentioned like Cloudera, Tableau, and Qlik. It is also possible to access services provided by Amazon and Microsoft.

Alteryx Designer tool provides also spatial information - e.g. the location of purchase with heat maps – and prepares the data for the tools having good visualization capabilities, like the ones Tableau and Qlik are offering. In addition, R statistical package is available as part of the processing scheme with possibility for own routines. The text analytics is possible by connecting to 3rd party tools like Microsoft (utilizing Cortana). Overall the prices are quite expensive, and spatial package is priced separately in the offering.

As a summary, Alteryx itself does not provide any own processing asset for text analytics, but they can be accessed from a 3rd party source. The challenges related to multi-linguistic support, qualitative analytics of unstructured data and multi-media formats remain to be solved by some other means than the Alteryx product offering. Getting all this information required couple of web conference meetings with the sales representative and some trials with the Alteryx Designer tool.

5.3.4 Microsoft Azure cloud services

Azure Machine Learning Studio is the product related to analytics. It is part of the Microsoft Cloud services, branded as Azure. Standard workspace is provided with $10/month and Azure subscription. R language and Python scripts can be integrated into the processing chain. At first glance, it may not support unstructured data, and it seems to be geared more towards prediction of factors affecting sales than natural language processing. This may not be the full picture, but it is not studied more in this paper.

5.3.5 Amazon machine learning

Amazon machine learning can help to process unstructured text (Amazon, 2016). It classifies product reviews as positive, negative or neutral. The pricing is based on the used

time of the service. Data need to be in Amazon's data store (S3, Redshift, RDS). Models are created, fine-tuned and then predictions are made. It is targeted to boost sales of the customers – like analysing information and making predictions affecting to sales or other factors related to it.

5.3.6 Some other commercial solutions for data analytics

SAS text analytics

SAS is having a bunch of tools related to predictive analytics, text mining and analysing unstructured contents (SAS, 2016). SAS Text Miner is said to provide "a rich suite of linguistic and analytical modelling tools specifically developed for discovering and extracting knowledge from collections of text content". Getting more detailed information on SAS Text Miner one need to request pricing or get personal contact to sales person.

There are many other companies that focus on data analytics and predictive analytics for sales purposes. Text analytics is usually mentioned as a part of the offering in the following list of tool providers. These were reviewed shortly during the investigation process: Hortonworks, Fico, Cisco, Angoss, Verint, Dell and Oracle.

5.4 Free software solutions for analytics

There were only few tools which were free software and yet having capability for analytics processing. The most promising ones were RapidMiner and Knime.

5.4.1 RapidMiner Studio

RapidMiner Studio if free up to 10 000 data rows, and for educational purposes it is totally free. From 10 000 rows of data up to 100 000 rows, it costs $2500 yearly when used for commercial purposes. Data needs to be typically in a structured format (excel, database formats, other analytics tools like SAS, SPSS), but also cloud sources like Twitter can be used as an input. It is possible to create different filters and views. Extensions to integrate Python and R project are supported.

In most examples, the data is like in a typical excel or database format. The tool is open source, but a license is needed for the commercial use. It is possible to use Hadoop servers for big data or high-performance computing. Quite long list of partners listed, looks like

a vibrant community. It is Code-free, i.e. no manual coding needed for data mining. Still it does not seem to work with unstructured text in the trial version (excel file or similar needed), but text processing extension is available in the RapidMiner marketplace. There was nothing mentioned about other languages than English.

5.4.2 Knime Analytics Platform

Knime is licensed under the GNU General Public License (GPL), version 3. So, it is really for free (Knime, 2017). It has a long history starting from 2008. It runs also in Microsoft Azure cloud. Java, R and Python coded methods can be added. For structured data analytics of statistical nature this looks a good tool, but it does not seem to support natively unstructured text.

There are many extensions and processing routines to KNIME to install and explore - from a complete R integration to such advanced topics as text processing and network mining. The user needs to develop the desired processing chain from the set of routines available. There seems to be quite active community of developers around the tool.

There is also extension for text processing, image processing (not contents nor sentiments) and social media including Twitter and APIs for Google Analytics and Google API connector. If the Google APIs mentioned in chapter 5.3.2 are accessible, Knime could be a complementary tool for Google analytic services.

5.5 Smaller companies with 'natural language first' approach

It is essential to find a solution which can detect qualitative elements and sentiments from unstructured data. Feelings and emotions are essential part of the feedback. This requires a deep understanding on the meaning of the language. Many solutions have been developed exactly for this purpose or having it as one of the main features. The algorithms for NLP are proprietary and secret for each company and the offered tool. The competition takes place on how well the algorithms perform in practice.

5.5.1 text2data

text2data is a name of the company and the tool (Text2data, 2017). It is developed for analysing unstructured data. Only English is supported, but there is an option to use the

Translation API provided by Google to convert other languages into English. The tool can be easily experimented. Here is an example what kind of word summary the tool can produce.

Figure 4. Word analysis done with text2data tool on the feedback campaign of Jaxber itself.

The size of the word describes how often it occurs in the text. The colour expresses positive or negative sentiment – red is negative and green is positive. As an example, the word important occurs quite often in the positive context.

The tool is affordable to use, and the charge is based on the usage or there are different monthly payment schemes. The textual units are stored e.g. into an excel with an add-in extension of text2data. The tool was experimented this way with the 356 textual cells, where each cell contained one feedback of a user of Jaxber. The most interesting part of the analysis was the sentiment analysis. Each text cell was evaluated separately by the tool. The next figure shows an extract of the output.

I think that this app must be modificated and have better design	positive	0,65448442	STRONG SUBJECTIVE
As I said before, I think it is a beta-version of app and that's why I have problems with bugs, uploading information and sometimes it's just end work by itself.	negative	-0,2495	WEAK SUBJECTIVE
In my opinion the most important value in this app that you are free in choosing in which way you want to answer. It can be text/video/sound or image.	positive	0,983178493	WEAK SUBJECTIVE
I think Jaxber is like an useful app for student's too ogives their feedback to lecturer. They can write their answers, film videos, share photos for each topic and teacher can easily check their answers and answer back, if it needs. But this app works not really good, I think it is beta-version and it must be finished.	positive	0,994040037	WEAK SUBJECTIVE
Without Jaxber, I could keep my learning diary in a Word document and upload it to Optima if needed. However, the minus point is I cannot view others' answer and receive comments from them as well	negative	-0,606493878	WEAK SUBJECTIVE

Figure 5. An extract of the output created by text2data regarding feedback collected with Jaxber

While looking at the sentiments created by the tool, it was not easy to always agree with the result. Personally, I agreed only in 27 cases out of 50 on the score given by the tool. The tool tends to give too positive scores. There were also opposite examples. The following was estimated erroneously as negative with a score of -0.214:

> *The most important element is Jaxber makes it easier to store the data in its own separate place and the data can be easily found from separate folders.*

The tool is easy to use. and it is interesting particularly regarding its capability for sentiment analysis. It could complement other tools, but is it reliable enough in its verdict?

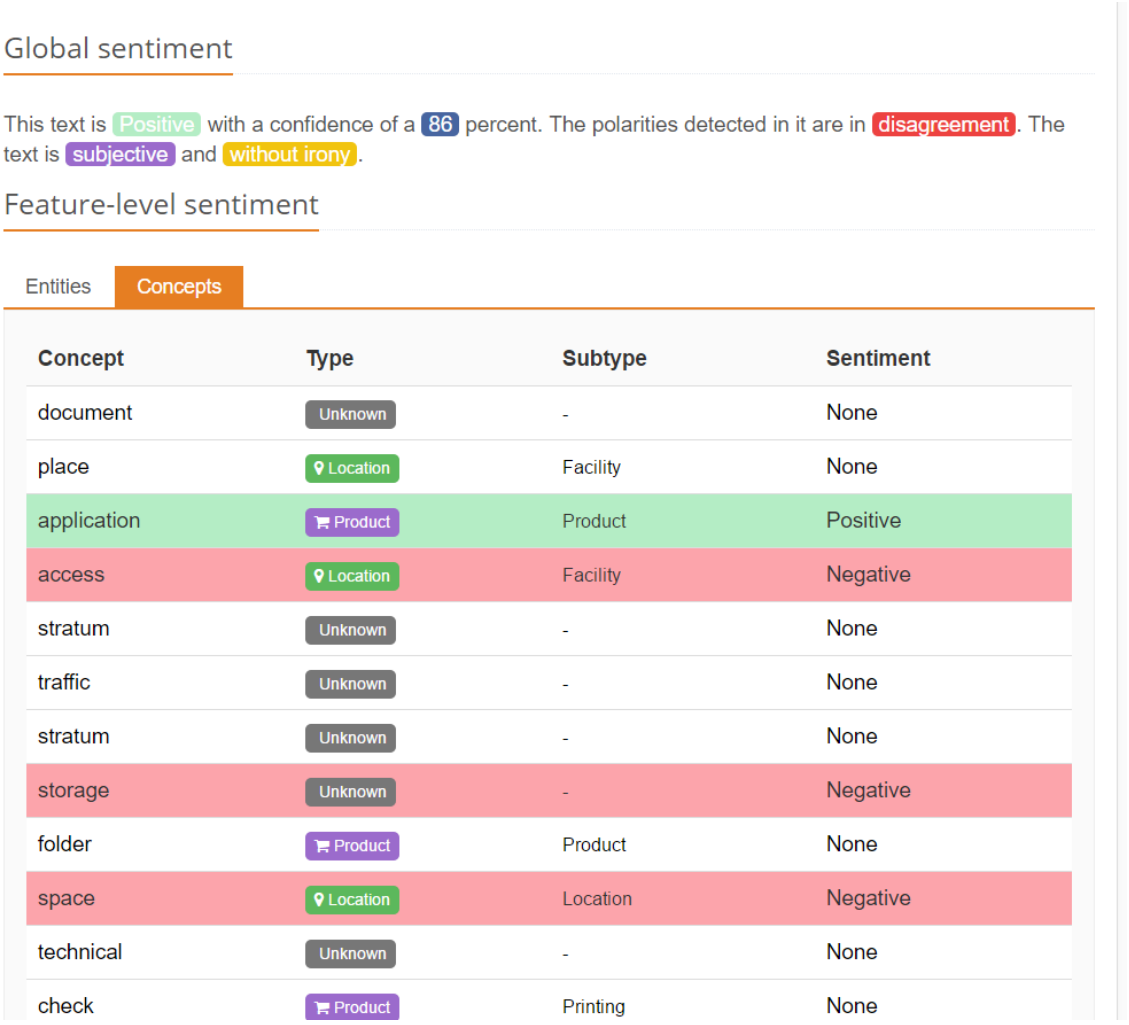
There are many scientific papers written in sentiment detection accuracy (Lin & Hu, 2009). The challenges relate to topics like: a) whether it is supervised learning or not b) is the corpora ("dictionary") known c) is there mechanism to convert negative words to positive when there is 'not' in the sentence? In the case of text2data it seems evitable that it does not convert negative words into positive ones (or vice versa) in sentences where structure is more complicated, like the first one in Figure 6.

5.5.2 MeaningCloud

MeaningCloud is quite similar tool as text2data for the same purpose. It is providing a cloud-based solution for automatic text analytics without any hassle of manual coding (MeaningCloud, 2017). The software is targeted particularly for analysing unstructured

customer feedback, and it supports Spanish, English, French, Portuguese, Catalan and Italian.

It is easy to try it with your own text up to 500 words. A quick trial demonstrates its capability to detect entities and concepts. It does classification and sentiment analysis – all that within few seconds for the own text content in the trial. The pricing is monthly-based, and if the data amount is very limited, it is for free. The price increases together with the amount of data. There are also APIs and web services available, which can be used from own applications. Here is a screenshot on the reporting style of sentiment analysis, where a sample of 500 words of Jaxber feedback was used as an input.



Global sentiment

This text is Positive with a confidence of a 86 percent. The polarities detected in it are in disagreement. The text is subjective and without irony.

Feature-level sentiment

Entities | Concepts

Concept	Type	Subtype	Sentiment
document	Unknown	-	None
place	Location	Facility	None
application	Product	Product	Positive
access	Location	Facility	Negative
stratum	Unknown	-	None
traffic	Unknown	-	None
stratum	Unknown	-	None
storage	Unknown	-	Negative
folder	Product	Product	None
space	Location	Location	Negative
technical	Unknown	-	None
check	Product	Printing	None

FIGURE 6. Sentiment analysis from a piece of Sodexo data performed by MeaningCloud.

Global sentiment part is quite self-explanatory summary of the analysis. Feature-level sentiment identifies different concepts, types, subtypes and their sentiments. E.g. ‘application’ is identified as a product which is regarded as positive.

There is an option to install MeaningCloud as an add-in into the Excel. There are selections for the following main functionalities: ‘Text Classification’, ‘Sentiment Analysis’, ‘Language Identification’, ‘Topics Extraction’ and ‘Text Clustering’.

A similar experiment was performed as with text2data. The textual content of the feedback concerning Jaxber was copied to the excel and add-in extension of MeaningCloud was activated to perform all the above-mentioned functionalities one at a time. The sentiment analysis gave better result compared to text2data. However, there were similar misinterpretations like in text2data, but not so much. Personally, I agreed in 31 cases out of 50 on the score given by the tool.

The tool is more cautious to give highly positively scores than text2data, and subjectively I agree more on the policy of MeaningCloud in that sense. The other features did not give much additional value without getting deeper into the features of MeaningCloud. It is possible to extend its capabilities by adding own dictionaries and defining own classification models.

5.5.3 Bitext

Bitext is claiming that their analysis methods are capable to achieve confidence level of over 90% for sentiments, categorization, entity extraction and concept extraction. They go linguistic first, aiming to understand sentence structure and layers of meaning for deeper text analysis. Support of over 20 languages is planned, and some 8-10 mentioned now as active including English, Spanish, French, German, Italy, Portuguese, Catalan and Dutch (Russian and Basque just coming). Customers include e.g. Intel and Movistar. They are providing API services for sentiment analysis API, Text categorization API, Entity Extraction API and Concept Extraction API.

It is easy to try out with some sample data. By registering, a trial period of 30-days of CX subscription was started, and the real data from Jaxber feedback was used to test the functionality. The generation of the report took over an hour with 350 entries of text, but it was worth it. It did a lot more than text2data and MeaningCloud. The following reports were created, and they can be viewed in a special dashboard:

- General dashboard on sentiment polarity
- Categorization
- Categories & Sentiment

The above reports were dynamic. User can add own filters for viewing the data. In Jaxber terms, the ‘Challenge’ field was used as a filter. E.g. user can choose ‘User experience’ as a filter, and then all the three views mentioned above are adjusted accordingly. You can look also individual topics that gets repeated like great, nice, reduce, crash, good – their sentiments and individual comments related. The dashboard provides filtering in three levels: type of comment filtering (Challenge), topic level and text level filtering.

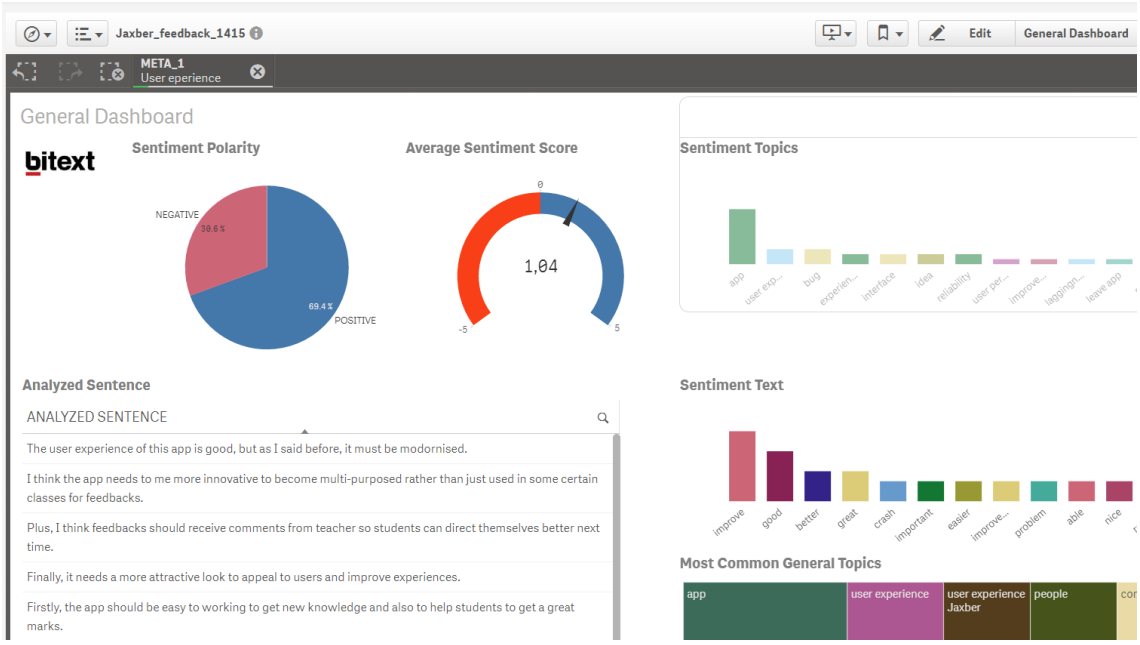


FIGURE 7. Example of a filtered view with Bitext, where user experience is selected.

Sentiment Topics – Such topics are shown which contain sentiments. E.g. above ‘App’, ‘user experience’ and ‘bug’ are such topics.

Sentiment Text - The most common words containing sentiments are shown. E.g. above ‘improve’, ‘good’ and ‘better’ are such words.

Analyzed Sentence – The list of such sentences which correspond the current selection of Sentiment Topics and Sentiment Text.

Average Sentiment Score – It shows whether the current selection of topics result in positive or negative sentiment. Blue is positive and red is negative.

If user selects say ‘improve’, the list of Sentiment Topics will dynamically change showing which topics relate to it. At the same time, Average Sentiment Score and list of analysed sentences are updated accordingly.

Thus, the real value of the figure is revealed when user can dynamically choose the topics shown, and see visually and textually immediately inter-dependencies. The tool creates automatically the topics and text items seen on the right. One can click each individual topic and item and the score and related sentences are shown, and vice versa – by clicking a sentence one can see to which topics and text items it is related to.

Browsing the dynamic report with multitude of ways enriches the way to inspect the original data. This is probably how the customer who ordered the campaign would like to review through the results. Let’s take another example with ‘Collective learning’ selected.

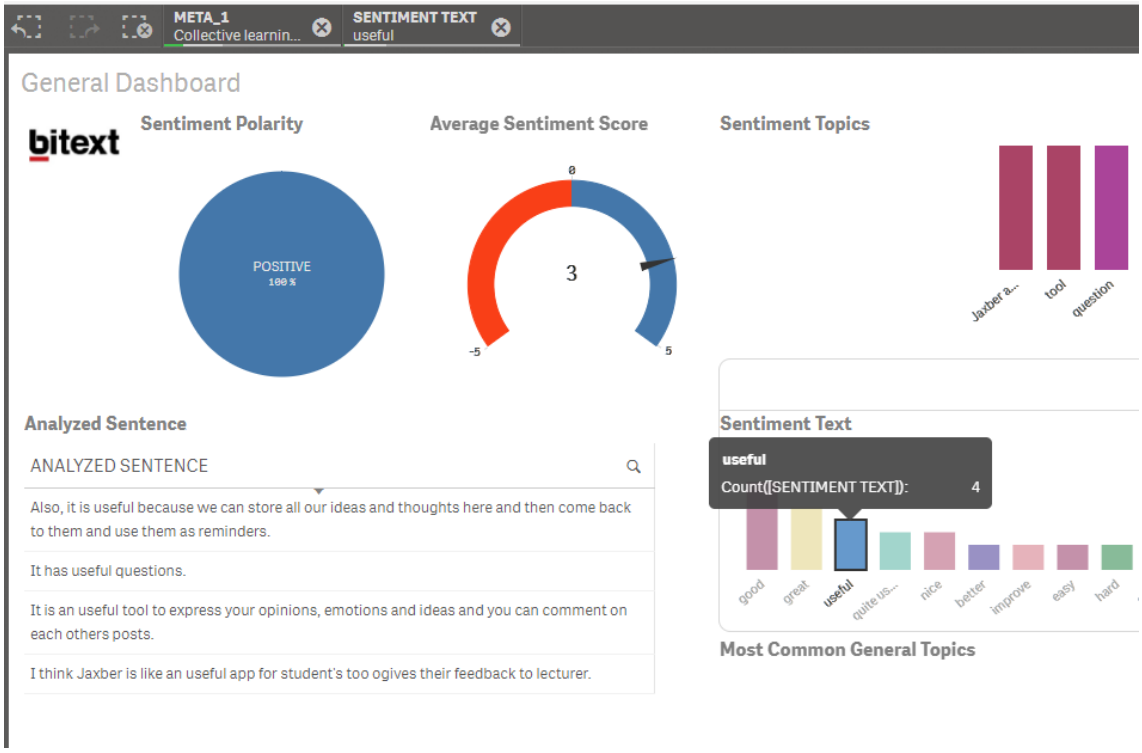


FIGURE 8. Example of a filtered view with Bitext, where ‘collective learning’ and ‘useful’ is selected.

The tool regard useful as a sentimental word, and it can be found in 4 sentences, which are shown on the left. The score is highly positive as being 3, and the word is relating to the following topics: Jaxber application, tool and question.

The price of monthly CX subscription is 599€ including maximum of 50MB data. The size of Jaxber data used as an example was 70kB. CX subscription is aimed for business users and marketers. Another option is to select Text Analysis API subscription which is focused on developing market research and analysing surveys. REST API, JSON output, asynchronous calls with https secured communication are technical features of it.

5.5.4 Summary on the tools with 'natural language first' approach

There are also many other companies. One of the them having quite promising NLP techniques in use was a Finland-based Etuma, but their pricing was overwhelmingly expensive. Also, a part of their offering is to collect the feedback, which makes it a competitor for Nestronite.

Out of the three solutions studied, Bitext proved to be the best one. It supports a wide range of languages; it does categorization and picks up topics of interest best of all. In addition, the dashboard gives value for the user with its data browsing capabilities. The option to use Text Analysis API instead of CX subscription could be studied more in detail.

6 Summary and comparison of the solutions

A comparison of the most potential alternatives is done in this chapter. The goals are first rephrased as more suitable for comparison. The list of alternatives is shortened significantly. Suitability of each alternative is assessed, and a proposal is made on how to continue to practical level, and what are the next actions towards a fully functional system.

6.1 Rephrasing goals / criteria for comparison

The original goals set in March 2016 are still valid. For comparison, it makes sense to shorten the list as follows:

1) How good analysis is provided (=G)
2) Support for multitude of languages (=L)
3) Readiness to convert video to text (=V)
4) Cost of the solution (=C)
5) Work effort to take the solution into use (=E)
6) How long it takes to take the solution into use (schedule) (=S)
7) Readiness to expand solution to operate in cloud, as "big data" (=B)

There were originally 16 goals, as listed in chapter 2.2.2. Each of them is part of the above list of 7 goals. So, basically no changes had happened in that sense. Regarding the priorities, goals L and V has become more important as thought early 2016. The Jaxber has developed towards multi-lingual direction rapidly and using videos as a mean for feedback enables it to differentiate from other similar tool providers. Goal G include factors like support for unstructured text, automated analytics and sentiment analysis, which were not specifically mentioned in the beginning.

6.2 Top-level comparison

The following four solutions stood out from the others:

- R Project
- IBM SPSS Modeler + Watson Speech to text
- Google Cloud Platform
- Bitext

The following table summarizes their score by 5 levels when compared to criteria in chapter 6.1. The letter chosen in chapter 6.1 is used followed by additional indication of --, -,

blanco, + and ++ signs. The signs are used as tokens of sentiments, ie. + means positive and e.g. in case of cost it is then cheap.

TABLE 1. Comparison of solution alternatives

Criteria	R	IBM	GCP	Bitext
How good analysis	G	**G**+	G-	**G**+
Support for languages	**L**+	L (~5)	L (3)	**L**+ (8)
Convert video to text *)	V--	V+ (6)	**V**++ (80)	V--
Cost of solution **)	**C**++	C--	C+	C-
Work effort	E--	E+	E	**E**++
When solution in production	S--	S+	S-	**S**++
Readiness for big data	B+	**B**++	**B**++	B

*) Criteria include number of languages supported.

**) Only immediate cost of purchasing included.

The selection of scores is partly subjective and based on a limited amount of information. However, the idea of the table is to show overall view on the comparison, and one should not pay too much attention on an individual score. The selection of scores is further justified textually for each alternative in the following chapters.

6.2.1 Suitability of R Project

There is a long learning curve to get knowledge and skills for using R Project in a productive manner. Even having worked with it full-day total of one year does not guarantee such competence level which would be required to code automated analysis from natural language. With 0-level as a starting point, it would really take a long time until the solution could be used in the production. As it is open source, there would not be any immediate cost of purchasing. How good would be the result, it depends on the amount of resources, their skill level for coding and the time used for coding. If the company has money to invest in in-house long-term development, this would be perhaps an option, but another intermediate solution would be needed meanwhile. Before doing so, it should be verified what is the capability of R libraries for sentiment analysis. It was mentioned as being a version from 2012 in one source, and if nobody develops it further, it may turn out to be insufficient.

6.2.2 Suitability of IBM SPSS Modeler and Watson Speech to Text

The clear benefit with IBM SPSS Modeler is that it is a tool which you can buy immediately and then start to learn what you can get out of it. The rapid experiment with it show that it can create some sensible output from unstructured text without any training data in less than hour of starting to use it without any previous knowledge. There are selections to indicate what kind of text is in question, which improves the outcome. E.g. the user can select that the text is concerning customer feedback. The price is high. In practice the premium version is needed to satisfy needs of Jaxber, and it costs €10904 per licence in annual basis. The Watson Speech to Text is cheap. I recommend to do experiments with it, and compare the accuracy to Google's Speech API. It is also worth trying how smoothly the output from Watson Speech to Text can be fed into the SPSS Modeler.

6.2.3 Suitability of Google Cloud Platform

Google does not provide any ready product for natural language processing, unlike IBM. However, an application could be built on top of the APIs listed in chapter 5.3.2. Whatever is selected as a tool, the Speech API is the most promising one for extracting information from the video files. The Vision API could be used for detecting information from the images collected with Jaxber. Currently the Natural Language API (Google, 2017) supports only 3 languages, and a brief experiment with it wasn't that promising.

6.2.4 Suitability of Bitext

With Bitext the user can get quite good overview on the input data without any big efforts. The dashboard provides a rich way for the customer to create various ways to make observations from the feedback data. It already supports 8 languages and there is more of them coming. The price is a bit expensive for monthly CX subscription. The example of Jaxber data amount was only 0,14% out of the maximum amount of the monthly data (50MB).

6.3 Hybrid model of tool selection

Considering all the formats of data which the Jaxber data is composed of, it is beneficial to use several tools to complement each other.

If the cost level is acceptable, for text processing the options are to take either IBM SPSS Modeler or Bitext in use. In annual level the comparable prices are 9900€ and 7200€, respectively. Regarding Bitext, some results can be achieved by using the free trial periods of 30 days.

For extracting information from the video files, the options are to use either Google Speech API or IBM Watson speech to text. Google has much better selection of languages, so whenever the language is not supported by IBM, the choice is Google. On the other hand, IBM is much cheaper, and it could be a good choice for using e.g. US English or UK English. While comparing with the same sample file of audio content, the IBM did slightly better, see Appendix 5. Note! This was just a single shot for comparison, so no conclusion can be drawn out of it. The input file was quite challenging, and some words were difficult to capture by any listener.

6.4 Conclusion and recommended actions

The hybrid model presented in previous chapter is recommended. For text processing, CX version of Bitext is recommended. Maybe cheaper price could be negotiated for smaller amount of monthly data as what is the standard offering. If Bitext does not support the language in which the feedback is given, then Google's Translation API could be tried even though this option is not verified in context of this thesis.

For Speech to text, the work with both options should be continued. As a priority, it should be started with Google Speech API due to much broader offering of languages. Comparisons regarding the speech quality should continue gradually with different audio files and speakers with different dialects.

As immediate actions, the following is recommended regarding Google Speech API:

- how to train context specific words, like Jaxber, there seems to be a way to do it. Google Speech API is talking about 'phrases' of type 'SpeechContext' in this context (Google-Cloud, 2017)
- how to process longer audio files than one minute

If all that turns out to be feasible, then the task is to plan an application which does the following processing steps for larger group of video files:

- convert first video file to flac audio file
- then convert it into Linear16 format, as descripted in chapter 5.3.2
- then send audio files into the cloud storage
- give command to Speech API to start the conversion
- receive results and organize the textual data as suitable for further processing by Bitext

The language options for the Google application are Python, JavaScript and C#. Python is a language which is used in Natural Language Toolkit (see chapter 5.2.3). There might be synergies in the future to utilize Python based libraries for text processing. So, it could be a good choice as a language for this task. It could be used also to pre-process the text converted from speech to text, if the raw data would be too difficult for Bitext to process.

The Google's Vision API can be used to interpret images. It can detect emotions in faces, and a broad set of objects. An alternative is also IBM's Visual Recognition, but it does

not detect text so well from the images, and it is not promoted to detect emotions in faces. So, Google's Vision API is recommended for image detection.

The following figure summarizes the proposal.

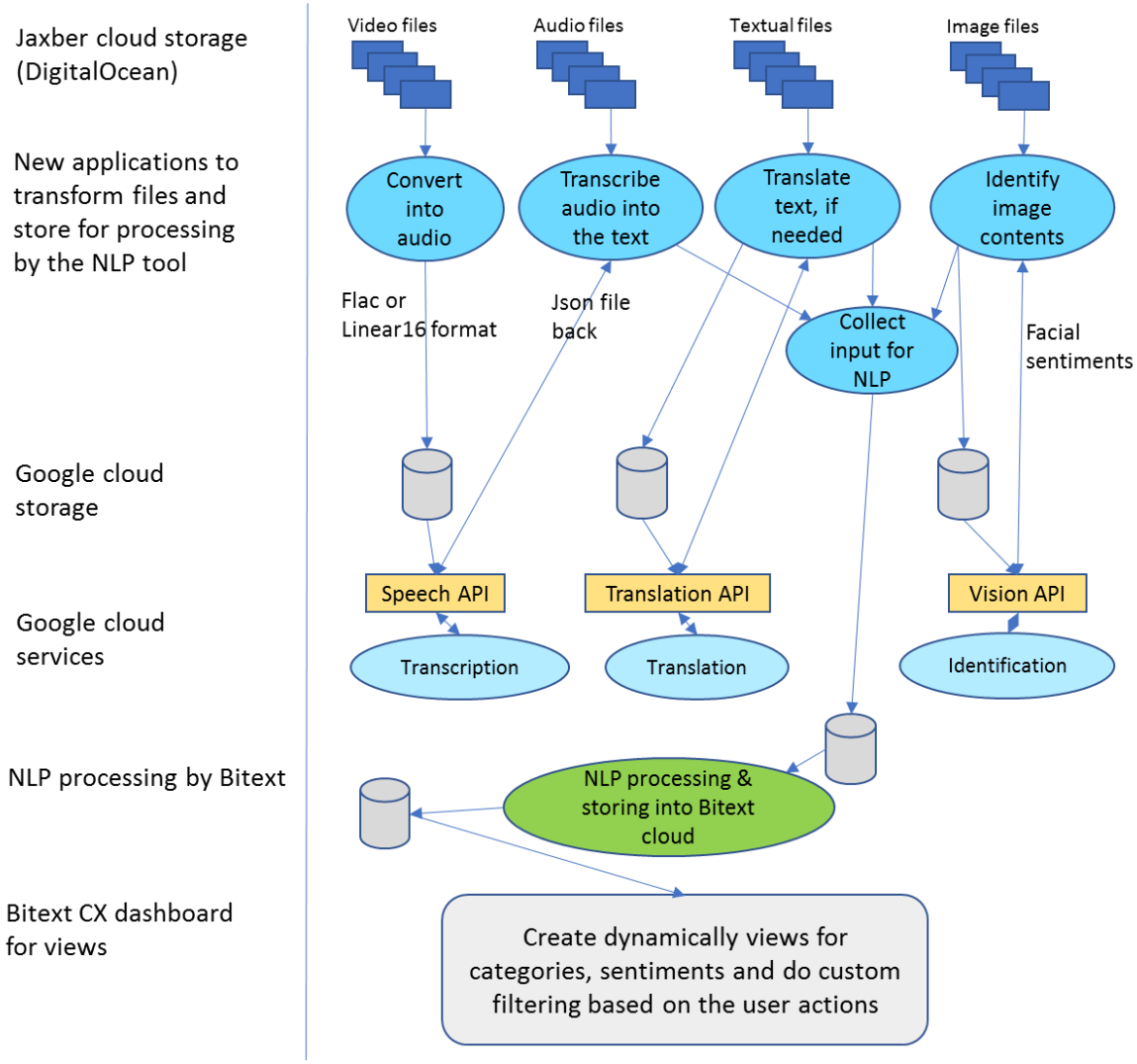


FIGURE 9. Conceptual illustration of the data flows and storages in the proposed solution.

If the solution is adopted, it should be considered to optimize the amount of different cloud storages.

7 DISCUSSION

This has been a true learning experience. In the beginning the goals appeared as almost overly ambitious to achieve. Even though the final breakthrough was not achieved, a lot of progress has happened, and this work proposes how to continue the work towards an ideal solution.

The focus was fully in the textual data in the beginning, and it was quite delightful to realize that it is not that far-fetched idea to extract also something out from video files.

The companies who have worked in natural language processing have typically strong academic background and experience of 10+ years in the area. Theoretical knowledge on statistical analysis is also beneficial. I recommend also to build such competencies as a long-term goal. Starting programming in such environment having built-in libraries for it is a good start. When level of theoretical knowledge and practical level programming skills increases, it provides basis for developing own solutions or modifying the source data. The motivation of modifying the source data could be data cleansing or e.g. pre-filtering it for specific needs of the customer.

REFERENCES

Amazon. *Amazon machine learning.* https://aws.amazon.com/machine-learning/

Apache. *Apache Hadoop Foundation*. http://www.apache.org/

Apache-OpenNLP. *Apache Open NLP*. http://opennlp.apache.org/documentation/1.6.0/manual/opennlp.html

Freeling. *Freeling open source language analysis tool*. http://nlp.lsi.upc.edu/freeling/

Google. *Google's Natural Language API.* https://cloud.google.com/natural-language/

Google-Cloud. *Google Cloud Platform, Speech API.* https://cloud.google.com/speech/

Hortonworks. *Source of HDFS architecture*. http://hortonworks.com/apache/hdfs/#section_2

IBM-Cloud. *IBM Speech to Text offering.* https://www.ibm.com/watson/developercloud/speech-to-text.html

IBM-SPSS. *IBM defining predictive analytics in context of their SPSS tool for statistical modelling and analytics.* http://www.spss.com.hk/corpinfo/predictive.htm

IBM-SPSSModeler. *IBM SPSS Modeler product pages with pricing.* https://www.ibm.com/marketplace/cloud/spss-modeler/purchase/fi/en-fi?S_TACT=000000OA&S_OFF_CD=10001871#product-header-top

IBM-Watson. *Announcement on IBM Watson-enabled data platform.* https://www-03.ibm.com/press/us/en/pressrelease/50846.wss

Knime. *Knime analytics solutions* . http://www.knime.org/knime-analytics-platform

Lin, C.;& Hu, Y. (2009). Joint Sentiment/Topic Model for Sentiment Analysis. *CIKM'09 Proceedings of the 18th ACM Conference on information and knowledge management.* Hong Kong.

LoopAI-Labs. *Common Machine Learning approaches*. http://www.loop.ai/machine-learning

MeaningCloud. (3. 1 2017). *MeaningCloud solution for NLP and sentiments*. https://www.meaningcloud.com/

Munezero, M.;Montero, C.;E., S.;& J., P. (2014). Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing*, 101-111.

Nestronite Oy. *Jaxber*. http://jaxber.com/en/

NLTK. *Natural Language Toolkit (NLTK).* http://www.nltk.org/

Pak, A.;& Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Proceedings of LREC.* Valletta.

Predictiveanalytics. *Predictive analytics web pages: www.predictiveanalyticstoday.com*. www.predictiveanalyticstoday.com

Predictiveanalytics, C. *Commercial software tools for qualitative data analysis*. http://www.predictiveanalyticstoday.com/top-qualitative-data-analysis-software/

Predictiveanalytics, F. (25. 1 2017). *Free software tools for qualitative data analysis*. http://www.predictiveanalyticstoday.com/top-free-qualitative-data-analysis-software/

Predictiveanalytics, *Commercial software tools for qualitative data analytics*. http://www.predictiveanalyticstoday.com/top-software-for-text-analysis-text-mining-text-analytics/

Revolutionanalytics. *Revolution analytics, on the companies using R*. http://blog.revolutionanalytics.com/2014/05/companies-using-r-in-2014.html

R-Libraries. *Comprehensive R Archive Network. CRAN Task Views, i.e. domain areas.* https://cran.r-project.org/web/views/

R-Project. *R project home pages*. https://www.r-project.org/

R-Studio. *Basic text mining in R*. https://rstudio-pubs-static.s3.amazonaws.com/31867_8236987cf0a8444e962ccd2aec46d9c3.html

SAS. *SAS text analytics*. http://www.sas.com/fi_fi/software/analytics.html#text-analytics

Text2data. *text2data web pages*. http://www.text2data.org/

Wikipedia. *Wikipedia definition of Data cleansing*. https://en.wikipedia.org/wiki/Data_cleansing

APPENDICES

Appendix 1. R-project usage examples

- There are instructions available how to use text analysis with R (text mining and sentiment analysis): https://www.r-bloggers.com/intro-to-text-analysis-with-r/.
- Instructions to start using RStudio with some most common commands to get started: https://eight2late.wordpress.com/2015/05/27/a-gentle-introduction-to-text-mining-using-r/

- Text mining of Twitter data with word cloud: http://www.rdatamining.com/examples/text-mining

Appendix 2. Text processing methods

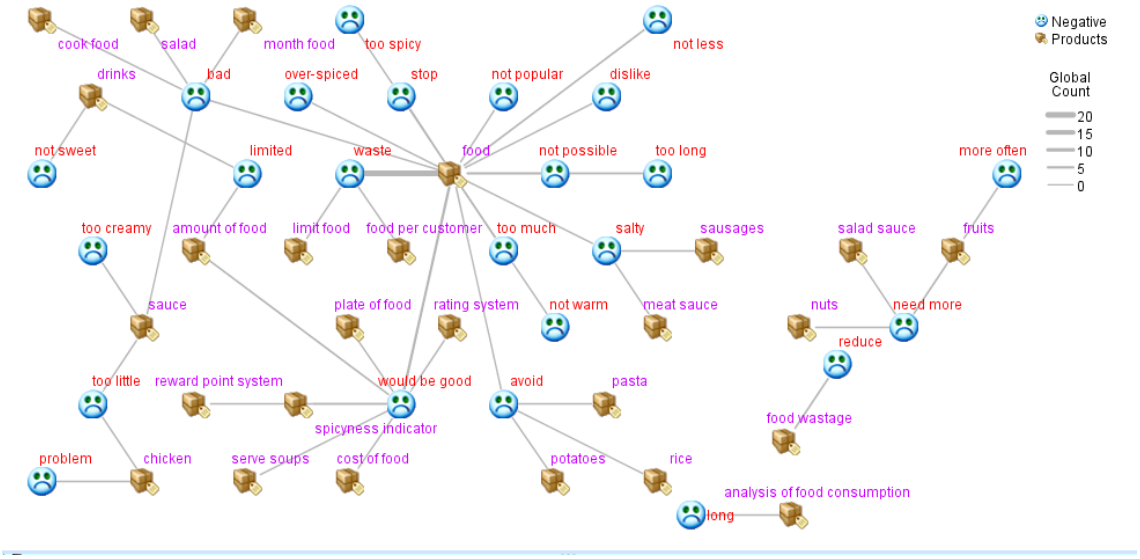
Name of the method	**Explanation**
Tokenization	Splitting text into words and sentences
Sentiment analysis	Extracting words or phrases triggered by the feelings and emotions. Typically, assessment done regarding if they are positive or negative, but other emotions possible too.
Part of speech tagging	Identifying the structure of the text and assigning each word into the corresponding grammatical category.
Categorization and classification	Typically, it is a question of pre-defined categories (sub-populations) having commonalities. Classification attempts to detect for each object into which category it belongs to. All this involves supervised or semi-supervised machine-learning algorithm which is trained with some example data to make decisions for classification.
Clustering	Discovering relevant topics and finding similarities, and then grouping them. Basically, similar task as categorization and classification, but as an unsupervised exercise without any training data.

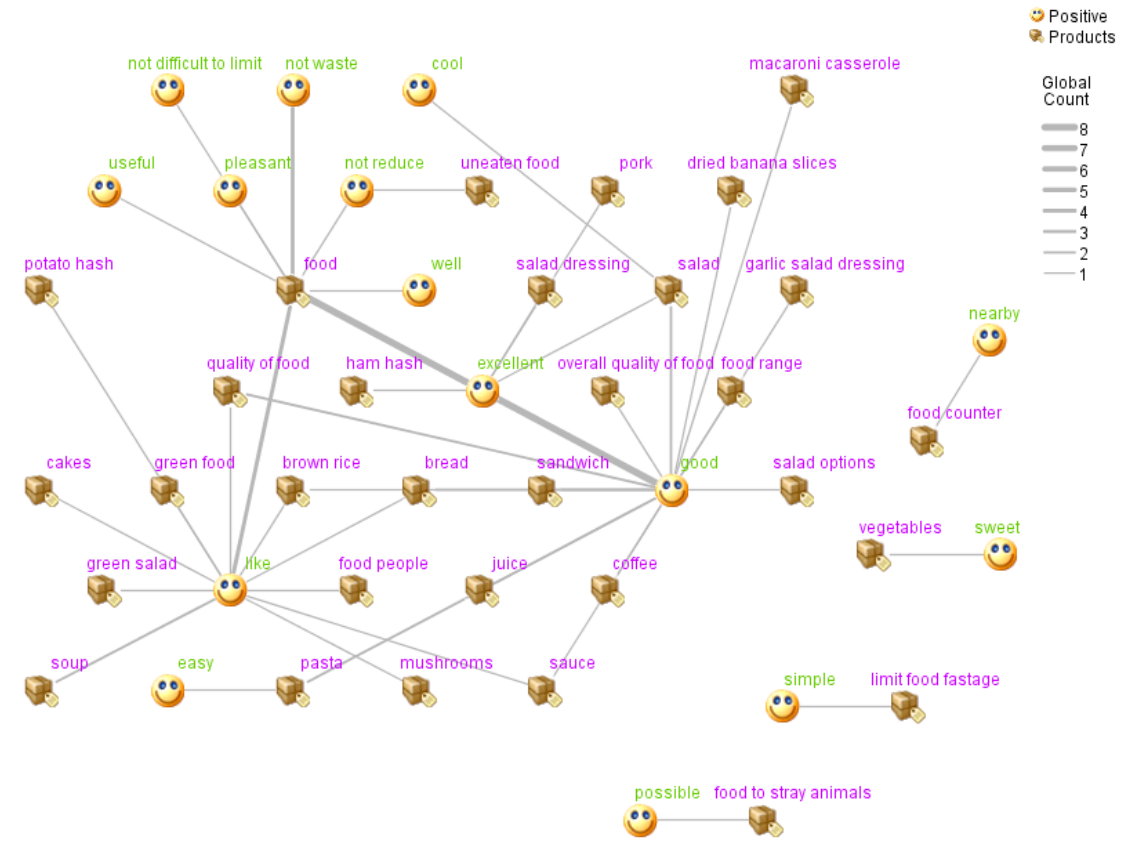
Appendix 3. Steps for experimenting Google's Speech API

The following steps were needed:

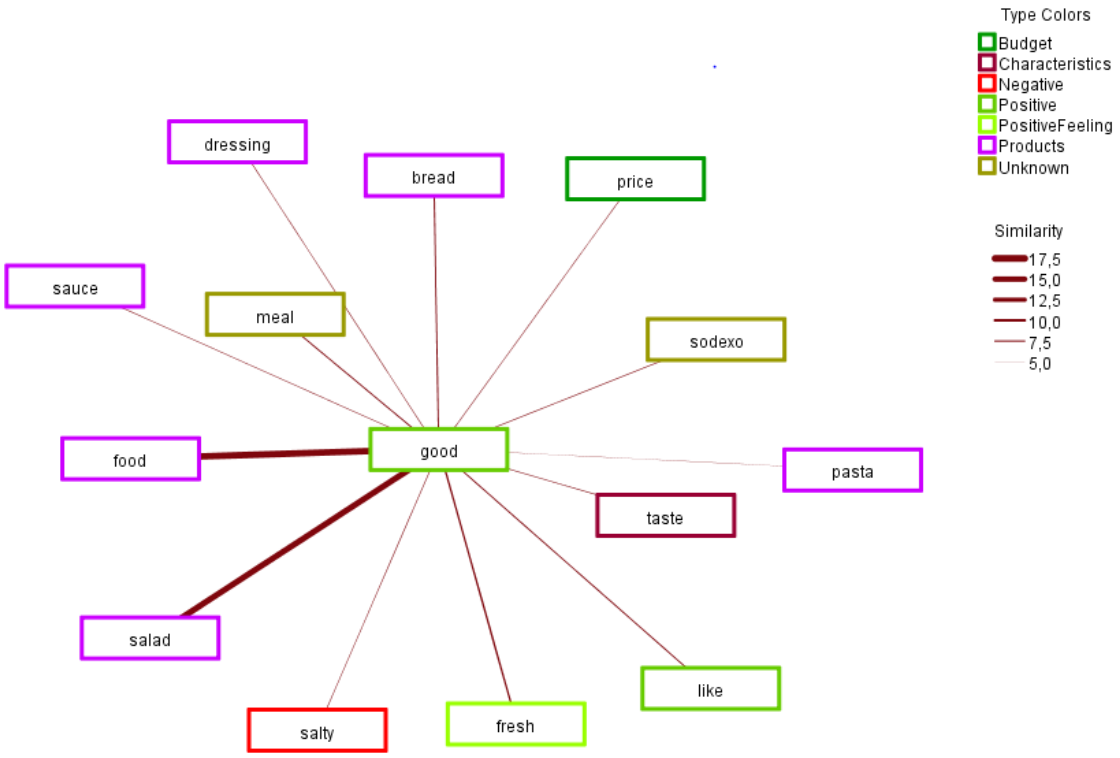
- Installing Google Cloud SDK with gcloud command line tool
- Registering as a user to the cloud
- Setting up your project with billing information
- Enabling the Cloud Speech API
- Setting up credentials for the Cloud Speech API
- Setting up a service account to authorize Cloud Speech API service
- Taking Cloud Storage into use
    - Creation of a bucket (a container or directory for files)
    - Copying audio source file into a bucket
- Creation of JSON request file, where reference to the audio source file
- Authentication of the service account
- Obtaining authorization token
- And finally sending the request with 'curl' command line tool (cURL) where the following parameters were passed:
    - json file referring to speech source file
    - credentials for authorization
    - speech:syncrecognize request

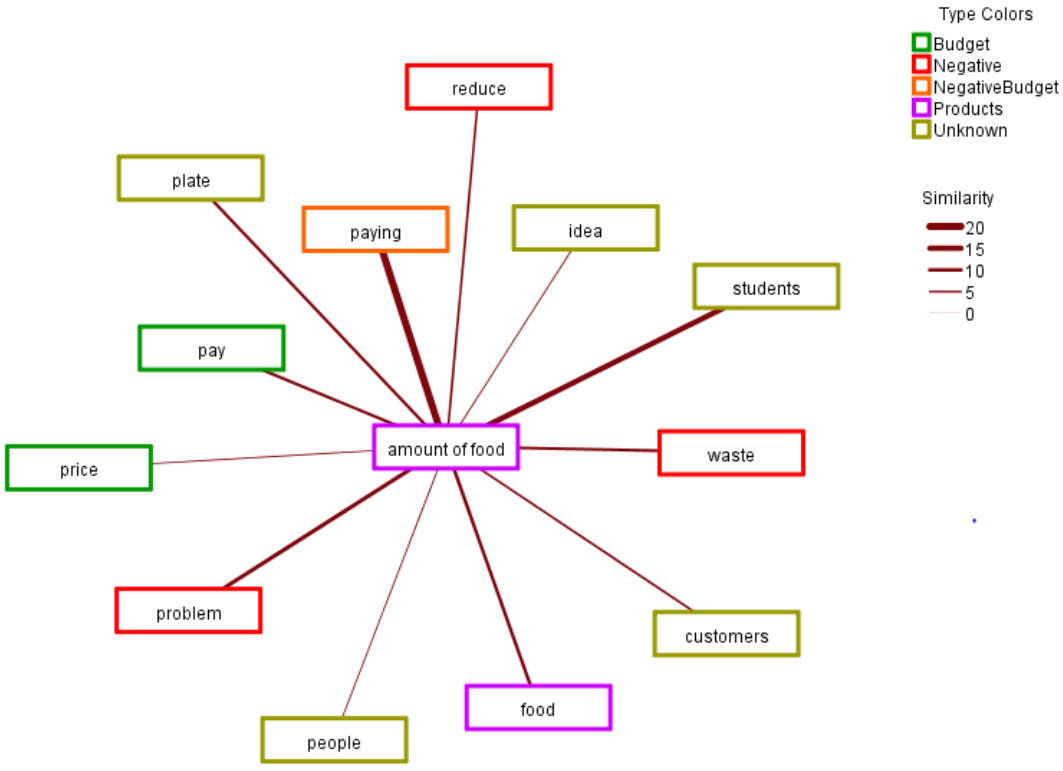
Appendix 4. Views generated with IBM SPSS Modeler on Sodexo feedback



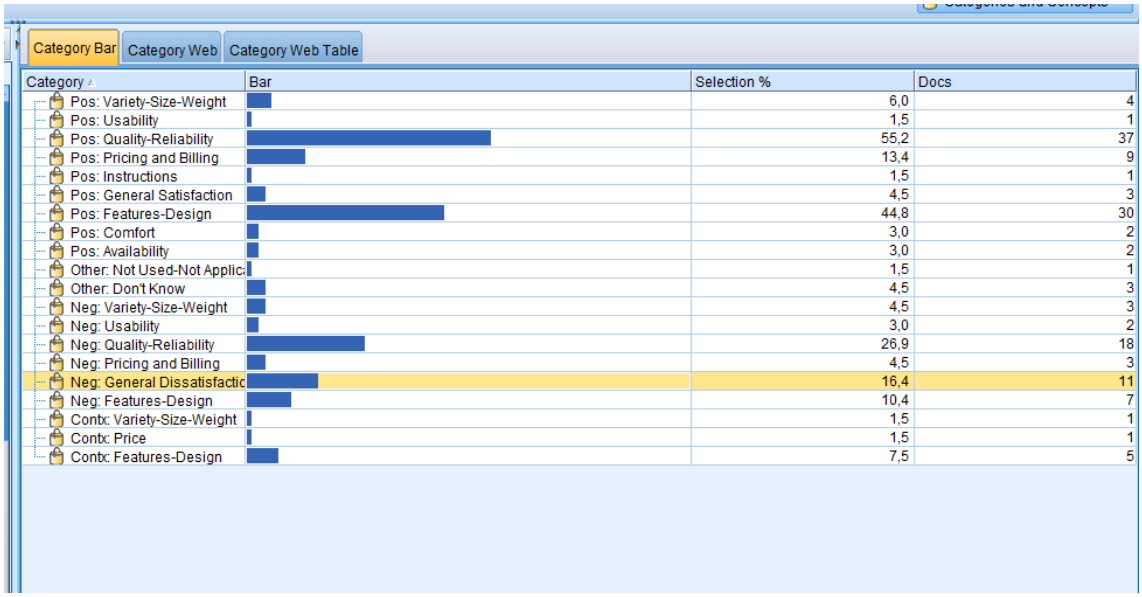


Positive and negative aspects related to catering services are expressed. The thickness of the lines describes how often topics relate to each other. Sentiments are expressed with emoji.





E.g. salad has been good quite often. There is strong relation in-between amount of food and paying.



Category	Bar	Selection %	Docs
Pos: Variety-Size-Weight		6,0	4
Pos: Usability		1,5	1
Pos: Quality-Reliability		55,2	37
Pos: Pricing and Billing		13,4	9
Pos: Instructions		1,5	1
Pos: General Satisfaction		4,5	3
Pos: Features-Design		44,8	30
Pos: Comfort		3,0	2
Pos: Availability		3,0	2
Other: Not Used-Not Applic		1,5	1
Other: Don't Know		4,5	3
Neg: Variety-Size-Weight		4,5	3
Neg: Usability		3,0	2
Neg: Quality-Reliability		26,9	18
Neg: Pricing and Billing		4,5	3
Neg: General Dissatisfactic		16,4	11
Neg: Features-Design		10,4	7
Contx: Variety-Size-Weight		1,5	1
Contx: Price		1,5	1
Contx: Features-Design		7,5	5

There are clearly more positive comments on the quality of the food than negative comments. Some general dissatisfaction exists. The user could see the specific comments from the tool better.

Appendix 5. A comparison of a speech content extracted from a video, with challenging accent

Original (as I hear it)	IBM	Google
Ok, these are our thoughts about Jaxber. Uhmm, we divided the whole topic into technical issues, missing features, aesthetics, value elements, engagement, missing features and alternatives.	Right in our own thoughts about checks for. Uhm we want to hold on to it in technical issues missing pictures. The next. Well you elements. Engagement missing features and alternatives.	OK Google Talk plug-in take me to Sydney beaches aesthetic value element engagement missing features and alternatives
In the center you can see our Jaxber thoughts. And we thought about general issues like things we like, we don't like. Ahm, first positive point there is a range of educational opportunities. You always connect with the students, students have to uhmm, think about the, uhmm, stuff they learned in class.	Some in the center in order to export. Um we just for a thought about the general issues with expertise in like the things we don't like. Uhm. First of all the point is the other range of educational opportunities you're always connected with the students students have to Rome.	in the centre you can see your expertise and we just thought I thought about the general issues with texts pretty things we like the things we don't like her suppose that point is there is a Range Rover educational opportunities you always connect with the student student have to think about the stuff they learnt in class
Uhmm, but also we had some negative points like the privacy issues. We don't know where our data is getting. There are pictures taken in class that are shared with a range number of people who we don't know who can see them.	Uhm but also we had a negative points like privacy issues. You don't know what it is getting. And their picture taken because they're here with a range of people we don't know who.	um but also we had some negative comments like the privacy issues we don't know where all data is getting their picture taken
Uhmm, students are forced to have smartphones and to download their app. There's no dexter alternative, and it is kind of maybe unfair for more pruce students.	Stephen uhm students are forced to have smartphones and to download the original text of alternative so. Kind of unfair for more for students.	
Uhmm, yeah there's good for storing information and sharing others, and, yeah.	Uhm. Yeah there's a scooter for storing information and sharing with others. And. Yeah.	
So, one of the, now maybe to the missing features. Uhmm, we said remin, reminders and notifications. If there are deadlines, we don't miss them. And chat options, where we figured out, especially with group works. It's great legs that you get contact in with students you never talked to before, and to find them to do organized group works.	So what the now maybe to the mission missing features. Uhm. Mister in remote locations of the deadlines that we don't miss them. And what chance option where we figured out especially with words. Uhm it's great lex that you can get in the persons and never talked to before and find them to organise a group first.	
Uhmm, multi device cloud that you can use it from the desktop. And then the Optima integration, so that you don't have three or four different, uhmm, places where you store data and share data, so maybe if Jaxber would be connected with …… would be nice.	Uhm advice cloud the UK and the us from the school. And then they often my integration so that you don't have a three or four different uhm. Places where you store data uh injured in maybe if checks for would be connected with the twenty or so yeah.	

Appendix 6. How to access Google's Speech API

json message:

```
{
  'config': {
      'encoding':'LINEAR16',
      'sampleRate': 16000,
      'languageCode': 'en-GB'
  },
  'audio': {
      'uri':'gs://pr1-data-bucket/kir-an55.raw'
            }
}
```

Note! The language code above is 'en-GB'. Of course, it refers to British accent. It proved to provide much better result for a Finnish speaker than using 'en-US', which refers to American accent.

A command requesting processing in Speech API:

```
curl --libcurl -k -4  -H  "Content-Type: applica-
tion/json"      -H      "Authorization:      Bearer
ya29.El_QA_41sCeNqlzdGtnNDbPZ2a0zvZdHtW2hOWM1jcTW
veZP1HtIJhW2LZT9NggprEy1KZeiBxgIhAddYGIJh5gbGk-
WVXoFnnZILP8S-FkhC729HADKSczWEWgggMPQ2w"
https://speech.googleapis.com/v1beta1/speech:syn-
crecognize -d @sync-request-latest.json
```