Developing and implementing a tool that combines and enhances current text analysis methods for persona development

Manuel Joana de Sousa Prata

Thesis to obtain the Master of Science Degree in

Industrial Management and Engineering

Supervisor(s): Ana Catarina Lopes Vieira Godinho de Matos
Jörg Michael Delhaes

Examination Committee

Chairperson: Prof. Miguel Simões Torres Preto
Supervisor: Prof. Ana Catarina Lopes Vieira Godinho de Matos
Member of the Committee: Prof. João Fernando Cardoso Silva Sequeira

December 2020

Declaração

Declaro que o presente documento é um trabalho original da minha autoria e que cumpre todos os requisitos do Código de Conduta e Boas Práticas da Universidade de Lisboa.

Declaration

I declare that this document is an original work of my own authorship and that it fulfils all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Resumo

Neste trabalho uma metodologia para o desenvolvimento de personas é criada e implementada. Esta metodologia faz uso de várias técnicas de análise de texto computacional, foi depois aplicada a um caso de estudo em que personas foram desenvolvidas, com base nas responsas a um questionário de resposta aberta.

O objectivo principal deste trabalho é o da criação de uma metodologia que combine ferramentas de análise de texto de modo a melhorar o processo de desenvolvimento de personas.

De modo a testar esta metodologia, foi desenvolvido um caso de estudo que consistiu em analisar as respostas a um questionário de resposta aberta a estudantes do Instituto Superior Técnico. As respostas foram recolhidas entre Maio e Julho de 2020. Findo este período de recolha de respostas, os dados foram analisados utilizando a linguagem de programação R. Sentiment Analysis, Clustering e Topic Modelling foram as principais técnicas aplicadas neste trabalho.

Com base nos resultados obtidos bem como numa interpretação manual dos documentos mais relevantes, foram criadas 3 personas: 'Scholar', 'Dominator', 'Stable Job'. Cada uma representa o perfil de um estudante hipotético com diferentes objectivos, crenças e motivações.

Os resultados obtidos estavam de acordo com as expectativas, como por exemplo, a confirmação de um víes relacionado com Sentiment Analysis reportado na literatura. Apesar de haver algumas limitações neste trabalho, estas podem ser ultrapassadas em trabalho futuro. Uma possível aplicação de técnicas mais avançadas de análise de texto ou um melhor processo de recolha de dados, deveriam melhorar os resultados.

Palavras-chave: Personas, Processamento de Linguagem, Questionário, Clusters

Abstract

In this work a methodology for persona development has been applied. This methodology makes use of multiple computational text analysis techniques, it was then applied to a case study in which personas were developed, based on the answers to an open ended questionnaire.

The overall objective of this work is to bring forth a methodology that combines current text analysis tools in order to improve persona development methods.

In order to test this methodology a case study consisting of an open ended questionnaire was presented to students from Instituto Superior Técnico. Data was collected from May to July of the year 2020. Following this data collection period, the data was analysed with the help of the computer programming language R, applying techniques from Natural Language Processing. Sentiment analysis, Clustering and Topic Modelling are the main techniques applied in this work.

Based on the results from these techniques, as well as an interpretation of the most relevant documents 3 personas were created: 'Scholar', 'Dominator', 'Stable Job'. Each represents an hypothetical student profile with different goals, beliefs and motivations.

The obtained results were according to expectations, such as the confirmation of a sentiment bias reported in literature. This work is a solid starting point towards improving the persona development methodology. Although there are some limitations in this present work, such shortcomings can be improved in future work. A possible application of more advanced text analysis techniques or a better data collection process are expected to improve the present work.

Keywords: Personas, Natural Language Processing, Questionnaire, Clusters

Contents

List of Tables

List of Figures

List of Abbreviations

NLP	Natural Language Processing
ML	Machine Learning
SVM	Support Vector Machines
NB	Naïve Bayes
ANN	Artificial Neural Networks
BOW	Bag-of-words
SA	Sentiment Analysis
IE	Information Extraction
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis

Chapter 1

Introduction

1.1 Motivation

This work sets out to create an efficient development of costumer personas. These were introduced by Cooper in his work (Cooper, 1999). Customer personas are used as a way to better understand and represent a user type, which uses a product in a similar way, due to shared goals, beliefs and needs (McGinn & Kotamraju, 2008).

Traditionally, personas were created with inputs from informal interviews to a small number of users. The interviews were analysed in order to find behavioural patterns and then these were grouped into personas (Goodwin, 2009). There were some drawbacks with this strategy, many of them dealing with the persona development step, such as lack of scalability, subjectivity and the fact that the process is slow (Miaskiewicz, Sumner, & Kozar, 2008). In the work of Miaskiewicz et al. (2008) it is stated that personas have their utility reduced by their development process, which is slow and complicated. These will be discussed in-depth in the Literature Review.

The analysis of answers to open questions (surveys, comments, etc.) has traditionally been done by human coders. In open-ended questions the respondents are asked to provide feedback using their own words, whereas in a close-ended questions there is a predetermined set of answers from which the respondents can choose. The use of human coders makes the process slow and prone to errors (Roberts et al., 2014). In this type of data one can find information that would otherwise not be collected and it is possible to find several different insights about the topic being analysed. The simple fact that these insights are not available from close-ended questions is a strong motivator for the application of NLP techniques to questionnaire data. Other big motivating factor is the complexity and time required to perform content analysis manually (Roberts et al., 2014). Instead of NLP a traditional content analysis could be applied. However this would be a much longer and more expensive process. Another disadvantage is that the researchers must first create categories and then assign the data to them, while when using Natural Language Processing and Machine Learning methods the categories can arise from data without being previously defined (Graneheim & Lundman, 2004). Furthermore the analysis by traditional methods is not very scalable, requiring more researchers or more hours to deal with an increasing volume of data, while computational methods are highly scalable by default (Silge, 2017).

In this work we develop personas from textual input using tools and techniques from the Natural Language Processing (NLP) research field. This field deals with the overall task of making computers capable of understanding free text (Feldman, 2006). With the help of these tools we can tackle some of the existing hurdles when dealing with answers to open-ended questions, namely the need for manual coding, the time needed to perform analysis and the problem of inter-rater reliability (Pietsch & Lessmann, 2018; Roberts et al., 2014; Schmidt, 2010).

This work is motivated by the fact that having a way to automatically analyse this information can overcome deficiencies in the current methods such as: having a faster time of analysis, consistency and high scalability of the process (Giorgetti, Prodanof, & Sebastiani, 2003; Schmidt, 2010). This allows for the creation of a persona development process based on the computational analysis of textual information addressing the issues previously mentioned.

Another motivating factor for this work are the advances made in the field of NLP (and related fields) that now allow for the development of better and easier to use tools (Mich, Franch, & Inverardi, 2004).

Furthermore, the following work has institutional as well as contextual relevance. In the February 2019 report by Comissão Análise do Modelo de Ensino e Práticas Pedagógicas (CAMEPP) [1] several changes are proposed and several flaws with the current teaching practices are exposed (Brogueira et al., 2019).

The present work seeks to provide a brief overview of how students perceive characteristics, it should be noted that there will not be a direct uncovering of flaws with the present work, instead it will allow us to understand how the students feel about the current state of affairs. One of the most direct measurements of this comes from asking the students how they feel prepared for their future in the workplace, since one of the flaws mentioned in the report is that despite amassing vast amounts of knowledge students do not feel ready to enter the workplace when they finish their course. It will also allow for the discovery of segments of students at IST. This information is useful when designing teaching practices. The present work accomplishes this through the development of personas based on a survey designed for this effect, this will be explained in the Research Methodology section.

Before proceeding with the rest of this section a mock persona will be introduced to help solidify the concept for the rest of the work.

This fictitious persona will represent a student as this is the type of persona we are going to be dealing with in this work. Beliefs and needs are aggregated into motivators, being the factors that make the student persona move forward. Here, the mock persona is presented.

Goals:	Motivators:	Sentiment:
goal a	motivator 1	Neutral
goal b	motivator 2	
	motivator 3	

Table 1.1: Mock persona.

[1] Available at https://conselhopedagogico.tecnico.ulisboa.pt/mepp-2122-modelo-de-ensino-praticas-pedagogicas-2122/documentos/

The bio for this persona is:

```
Mock is a -nd year student. He highly values motivators 1 to 3. He is neutral in sentiment.
He lives for achieving goals a and b .
```

Along with these elements each persona will have a fictitious student card. The card will follow the template presented in the figure below.

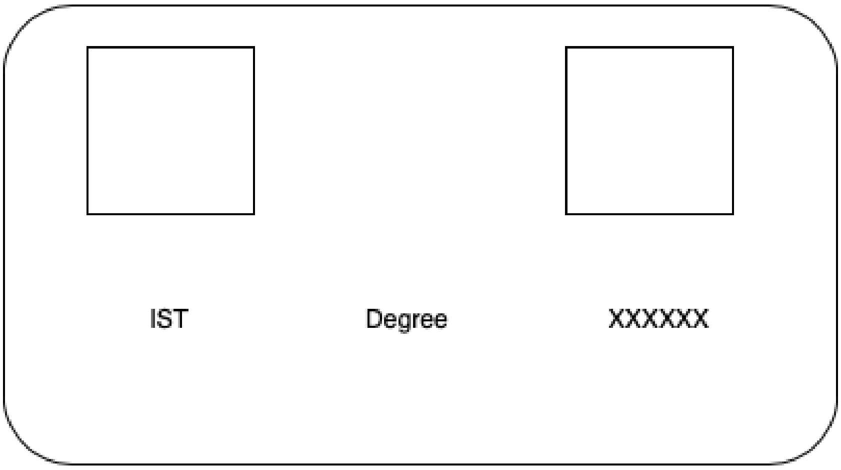


Figure 1.1: Mock persona student ID card.

Moreover, the fact that free text is a format in which there is a lot of stored information, makes this work even more valuable, as it provides a way to give structure to an unstructured format, making knowledge acquisition easier (Silge, 2017). These NLP techniques can be applied to problems other than developing personas, like the automatic grading of an essay, analysing user reviews and some even have applications in bioinformatics (Maalej, Kurtanović, Nabil, & Stanik, 2016; Rokade, Patil, Rajani, Revandkar, & Shedge, 2018; Zeng, Shi, Wu, & Hong, 2015). With this it can be seen that this work is in an area where there are multiple possible applications and possible research avenues, making it an even more interesting work.

1.2 Objectives

The present work has several objectives, in the following paragraphs they will be enumerated and briefly explained.

The overall objective of this work is to bring forth a methodology that combines and enhances current text analysis and persona development methods. The main objective here would be to develop a systematic persona development tool. The present work requires an analyst to perform the final step of creating and presenting the personas, meaning that this work focuses on preparatory data-analysis work to ease persona creation.

To meet the overall objective of this thesis, specific objectives were set. First a literature review was performed to analyse the existing options for textual analysis. In this work the R computing language is used as a main framework for the analysis. This is done since the language provides an adequate framework for the task at hand (Feinerer, Hornik, & Meyer, 2008). On top of this R is an open-source project. In this objective is included the selection of techniques for textual analysis. From the many techniques available only some are selected, since these are the ones relevant towards the goal of developing personas.

Following the selection of computing framework and techniques to use the present work focuses on developing and implementing the tool.

Finally, the second objective consisted of assessing the application of the proposed methodology in a case study. The tool was applied to a real dataset, consisting of answers to a questionnaire that was delivered to the student body of the University ”Instituto Superior Técnico” (Técnico/IST). The questionnaire was delivered using the Google Forms® platform. In order to reach a large number of students several student organisations were asked to mail the questionnaire link, which was available during a two week period.

Afterwards the results will be analysed and discussed. This will then be used to develop personas representative of the population under study, which are the final product of the analysis.

1.3 Outline

The document is structured in the following way: in chapter 2 the problem definition is presented. In chapter 3 a literature review is provided, regarding the techniques and methods used throughout the course of this work, namely NLP, Machine Learning and a section on personas and their development is provided. In chapter 4 the methodology for the work is detailed, namely what are the methods used to perform each task in the finalised tool. In chapter 5 the main results from technique are presented and analysed. In this chapter there is also a section detailing the case study that was performed as a methodological test for this work. Finally, in chapter 6 conclusions drawn from this study are presented, as well as the achievements of this work and what is left to be developed in the thesis. Chapter 6 also contains a section detailing a discussion regarding the current work. This organisation is depicted in Figure 1.2 .

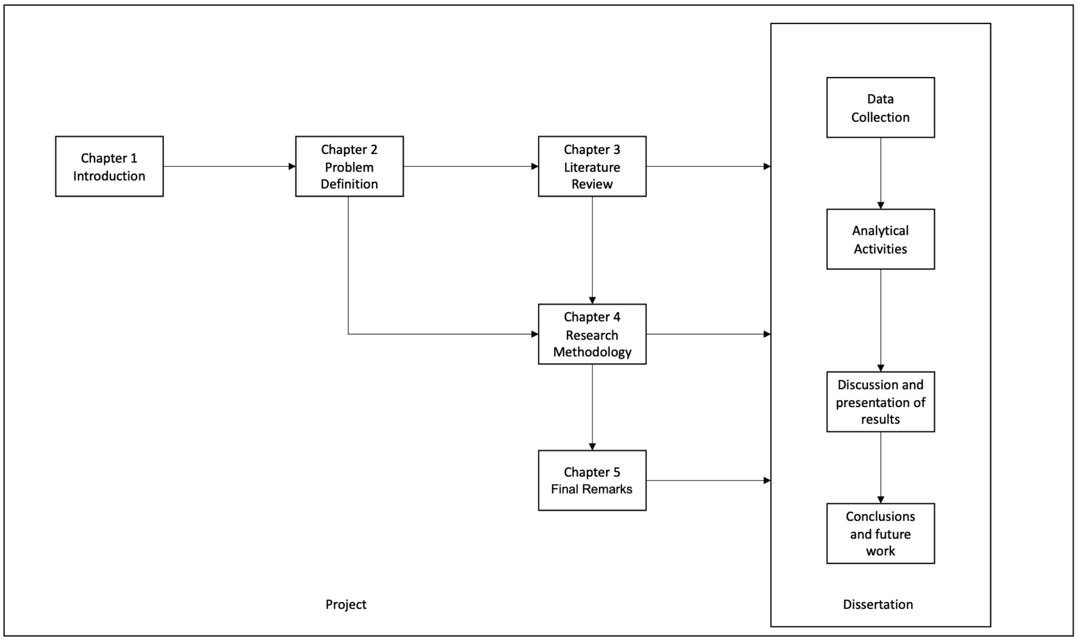


Figure 1.2: Proposed workflow, along with division of the project stage, and the thesis stage

Chapter	Dissertation activity
Introduction	Provides context and motivation
Problem Definition	Lists potential problem areas
Literature Review	Ascertain methods and techniques
Research Methodology	Detail methodological steps and software used
Final Remarks	Lists benefits and challenges

Table 1.2: Roadmap of dissertation activities

Table 1.2 also includes a roadmap containing the technical activities performed, roughly segmented according to the different chapters.

Chapter 2

Problem definition

The efficient creation of personas is a twofold problem. Traditional and some current methodologies for developing personas are based on the manual analysis of qualitative data such as interviews transcripts, direct observation notes and surveys. The goal is to identify behavioural patterns and turn them into a set of characterisations (Goodwin, 2009). This is essentially a manual clustering technique based on expert judgement. This is a slow and expensive process, that can greatly benefit from automation from NLP and ML knowledge areas (Miaskiewicz et al., 2008). Moreover, this process is not very easy to reproduce, with this also being a problem. More on these topics is discussed in the Literature Review.

Aligned with this goal, a methodology based on Natural Language Processing techniques was designed. This methodology is the engineering part of the problem tackled in this work. Several problems arose during this time, namely: problem area 1 is related with the computational analysis of free text; problem area 2 is related with proving that the presented idea works in a case study, and, finally problem area 3 that relates with good questionnaire design, which is needed in order to derive meaningful/expressive personas.

With these problem areas in mind, we can better understand why this problem also meets some challenges in areas other than engineering. The design of the questionnaire, its interpretation, the interpretation of collected data are all related with social sciences, as these are steps which intend to provide information relating the relationships and viewpoints of the respondents.

Problem area 1

In this work the main technical focus is the extraction of relevant information from the responses to an open-ended questionnaire. This information will then be used to accomplish the more case oriented goal of this work, the characterisation of students and subsequent representation of this characterisation as student personas.

Traditionally, the extraction of information from this type of data is a slow process, which requires a considerable amount of specialised human resources (Roberts et al., 2014). This creates a bottleneck in the acquisition of knowledge, especially because the type of information one can obtain through open-ended questions is not easily obtainable through closed questions. The answers to closed questions are much simpler to analyse, making it easier to automate said analysis (Fielding, Fielding, & Hughes, 2012).

The process of using humans to code the answers, used traditionally to analyse open-ended questions, creates a reliability problem: inter-coder reliability. This problem refers to the extent that two or more independent coders agree on the coding. This issue arises because despite having objective coding rules, the interpretation of the text might lead the raters to classify it differently, making the analysis less reliable (Lavrakas, 2008). Adding to this, coders are faced with very subjective decisions and arbitrary judgements, which may distort the original intent of the respondent. This is a serious issue, because one of the main advantages of asking open-ended questions is to elicit information about issues the respondents feels are relevant, in their own words (Looker, Denton, & Davis, 1989).

Furthermore, the time needed to reach an acceptable level of inter-coder reliability may make the systematic analysis of open-ended questions undesirable or even impossible (Mossholder, Settoon, Harris, & Armenakis, 1995).

Problem area 2

In order to work around the processing bottleneck generated by needing human coders to analyse answers to open-ended questions, this work will apply techniques from Natural Language Processing and Machine Learning. The goal of applying these techniques is to create an analysis pipeline for the analysis of open-ended questions, which is not heavily dependent on human labour, is easy to apply and can generate reliable analysis in a short span of time.

A case study is needed in order to verify if the proposed analysis chain is viable. Furthermore, using a case study is an effective way of investigating phenomena within their real life contexts (Eisenhardt, 1989).

Problem area 3

One should note that a problem faced by researchers when dealing with qualitative data is the lack of clearly formulated methods of analysis to deal with this type of data (Miles, 1979). This problem has been somewhat mitigated by the research community in the past years, with works focusing on the methodological aspects of qualitative data analysis (Brouse, 2002) (Burgess, 1994). Despite this, the process of analysing qualitative data is still very time-consuming, due to computational complexity (Burnard, Gill, Stewart, Treasure, & Chadwick, 2008).

Since open-ended questions are a very effective method for measuring important concerns for the respondents, this is a type of survey that can elicit various forms of useful information about organisations, institutions and even about public events (Geer, 1991).

Despite being simpler to analyse, close-ended questions have as a major drawback the fact that they only allow information to be elicited about a specific topic, whereas open-ended questions allow the respondents to use their own words. More importantly, they allow the respondents to mention topics relevant to them that might have been missed had the question been posed in a closed form (Engwall, 1983).

Among open-ended question there are three types, the technically open-ended questions, the apparently open-ended questions and the really open-ended questions: the really open-ended questions elicit specifications or reasonings from the respondents (Popping, 2015).

This work focuses on really open-ended questions. Due to their nature, these type of questions help us collect information which can be used to differentiate between elements of a population, such as preferences or sentiments regarding certain topics.

As the proposed work is methodological in nature, a case study which allows the testing of said methodology is required. In order to do so, and in-line with the recent report by Comissão Análise do Modelo de Ensino e Práticas Pedagógicas (Brogueira et al., 2019) a questionnaire was developed that elicits information about how the students feel about the current teaching practices and how prepared they think they are for the future as professionals.

Chapter 3

Literature review

3.1 Persona design

Personas, in their original conception, are groups of similar users of a product that have the same needs, goals and objectives. Personas help in the design process because they help the designer get a better perspective on the user of the product (Goodwin, 2009). Personas were introduced in the late 1990's by Alan Cooper as a design tool. Further refinements were made on the concept, originating Goal Directed Design (Cooper, 1999).

The personas used towards this goal are called customer personas. Customer personas are referred to as ”personas” from now on.

It is relevant to mention that there are other types of personas such as the future persona, which are personas that might exist in the future and can be used to complement a scenario (Fergnani, 2019).

This work will focus on customer personas as they are suitable for the proposed objective of characterising students . These personas serve the purpose of representing a subset of the users. Generally, a persona is made up of the following elements: fictional name, job titles and major responsibilities and some demographics. On top of these basic elements, personas also have information such as their goals, beliefs and environment (Pruitt & Adlin, 2005). This information about the goals and beliefs of students is what will allow the differentiation. Personas help overcoming biases and assumptions about the users being described (Miaskiewicz et al., 2008).

In this work, personas will be used as a way to communicate the findings from the analysis of the questionnaire by grouping students who have similar beliefs and expectations about their education, meaning that the personas will be developed as a final stage of the proposed methodology, being the main deliverable.

Previous works related to developed personas as a way to characterise students are rare, research around personas is more centred around Goal Directed Design rather than focusing on the more humanistic side of personas as an end itself (Cooper, 1999). Current research is more focused around using personas as a tool for the design of a product or service (Goodwin, 2009; McGinn & Kotamraju, 2008; Pruitt & Grudin, 2003). The creation of student personas can help get an enhanced view of the student population, ultimately allowing for a better design of learning experiences (Lilley, Pyper, & Attwood, 2012). In the work of Lilley et al. (2012) personas are developed for a specific group of students, those participating in distance learning activities. They used an hybrid methodology where personas were first developed ad-hoc, meaning that they were developed using information provided about the students, but not by the students themselves. These are used to develop criteria used to build the more detailed personas based on data provided by students (Pruitt & Adlin, 2005). The data was collected in two rounds, the first consisting of an online survey and the second consisting of interviews. The findings from the interviews were used to amalgamate the personas derived from the survey to a more manageable number of five personas (Lilley et al., 2012).

3.1.1 Persona Development Process

Customer personas are usually created by researching potential users, patterns of behaviour are discovered during research (Cooper, 1999). Following this research, experts comb through the data, which might be available in the form of questionnaires, focus group interviews or interviews, in order to find patterns. From these patterns the personas are constructed. Personas are representations of users with similar beliefs, attitudes, goals and needs (Pruitt & Grudin, 2003). This approach requires more than one expert in order to have some sort of validation of the generated personas. This is done by cross-checking the personas generated by each expert in the team (Goodwin, 2009).

This approach has some limitations, namely, the fact that it is an expensive and slow methodology to apply correctly (Hill et al., 2017). These limitations coupled with the fact that it is hard to validate personas and that they might be created from data that is not representative of the entire user base are strong motivators for the development of new methodologies that can tackle these deficiencies (Chapman & Milham, 2006). NLP techniques can help mitigate the concern of not having data representative of the entire user base, since in traditional methods the amount of data needed to do this is a hurdle, the analysis of this amount of data would be too long (Miaskiewicz et al., 2008). Computational methods can analyse large volumes of data in a short amount of time, given a reasonably large sample better representation is achieved.

Given these limitations new methodologies for developing personas started to emerge. The main difference is that these new methodologies use quantitative analytical techniques to explore the data, such as clustering and dimensionality reduction techniques. These have the advantage of being much faster and cheaper to employ, while also not being subject to biases from the research team developing the personas (Thoma & Williams, 2009). It is important to note that these quantitative methods can still be biased, depending on the sample of data chosen to develop the personas. It is possible that this is a major problem since validation is hard, as was mentioned beforehand.

The methodology from Thoma and Williams (2009) consist of heuristically combining data from several sources both qualitative and quantitative. The quantitative data was analysed using clustering and principal component analysis techniques. With these patterns the personas were developed and refined using the qualitative data.

In McGinn and Kotamraju (2008) a specialised survey was analysed using factor analysis. Factor analysis is a statistical method for describing the variability of correlated variables using a lower number of unobserved variables called factors (Yong & Pearce, 2013). These factors are then grouped. The survey consisted of 18 multiple choice questions that gathered domain-specific information as well as demographics. The patterns of behaviour were then used to drive persona development as suggested in (Cooper, 1999). This method allows the personas to arise from the data, meaning that it is not subject to bias from the researchers when generating the personas (Salminen, Jansen, An, Kwak, & gyo Jung, 2018).

In the work developed by Miaskiewicz et al. (2008) a methodology similar to the one that is developed in this work is explored. Latent Semantic Analysis is used in conjunction with clustering techniques to identify personas directly from textual data (Miaskiewicz et al., 2008). Their methodology consists of computing distance measures between the textual transcriptions of interviews and then using clustering techniques to group these. This study does not guarantee that the methodology can be generalized. Another issue stems from using cosine distance as a measure for similarity, as thresholds for high and low cosine distance measures are not well defined (Miaskiewicz et al., 2008). In this work a general semantic space is used meaning that domain specific words are not recognised properly. The persona narratives are still written by a human expert, which ultimately decides what is written in the narrative. This is in line with the use of computational text analysis techniques in this work.

The works discussed above are not an exhaustive list of all methods currently available for the development of personas, nonetheless they provide a view of different methodologies and ways to combine them into methodologies for data-driven persona development.

In the remainder of this chapter a brief overview is provided about which are the useful tools and methods to solve the addressed problems in this work. We shall also discuss the underlying assumptions made when applying each technique or model. Furthermore, the purpose of each task regarding the scope of this work will be discussed.

3.2 Computational text analysis

3.2.1 Machine Learning

Machine Learning (ML) is the study of computer algorithms that allow computer programs to automatically improve through experience (Mitchell, 1997). It is comprised by a set of methods that automatically detects patterns in data, and then use the uncovered patterns to predict future data, or for other kinds of decision-making under uncertainty (Murphy, 2012).

Machine learning techniques can be divided into supervised and unsupervised techniques, with the difference between them being that supervised learning techniques require that labeled data is available for training.

Supervised learning deals mainly with regression and classification tasks. These tasks are somewhat similar, but, in regression, the output is comprised of continuous values, whereas in classification the output takes the form of discrete values (class labels) (Marsland, 2014). Supervised learning techniques are sensitive to biased training data, replicating the bias in the classification task (Madhoushi, Hamdan, & Zainudin, 2015).

In unsupervised learning, the main tasks are clustering, discovering latent factors and discovering graph structure. Such algorithms first uncover a structure hidden in the data, and then exploit this in order to, for instance, organize it (a clustering algorithm) (Marsland, 2014). Unsupervised learning techniques are able to find patterns in data automatically, which is useful for finding connections in large volumes of data (Madhoushi et al., 2015). However, such techniques have the drawback of needing a larger volume of data when compared with supervised learning techniques(James, Witten, Hastie, & Tibshirani, 2014).

In the next subsections we will show typical examples of both learning strategies.

3.2.2 Supervised Learning

Support-Vector Machines

Support-Vector Machines (SVM) are a class of algorithms that implement the following idea: input vectors are non-linearly mapped to a high-dimensional feature space (this mapping is called kernel) and then a linear decision surface is constructed. The decision surface is the boundary that best separates the possible outputs. In this case the margin between support vectors is maximised (Cortes & Vapnik, 1995).

The following figure will illustrate the idea of support vectors and margins in a two dimensional space. The algorithm is very powerful, since decision surface properties ensure high generalisation. This is accomplished by maximising the margin between support vectors, in the transformed, high-dimensional space (Cortes & Vapnik, 1995).

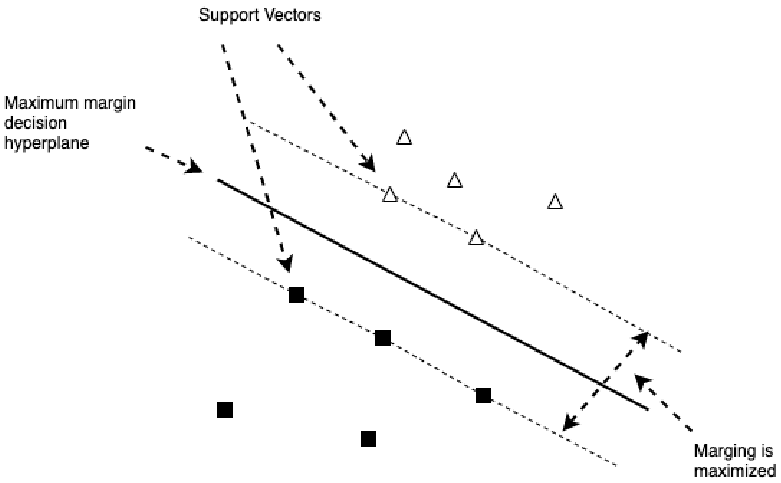


Figure 3.1: Support Vectors and margins. The maximum margin decision hyperplane is the decision surface. Figure inspired by (Manning et al., 2008).

In SVM, complexity is not affected by the number of features, so it deals well with high dimensional data and has good generalisation ability (Singh, Thakur, & Sharma, 2016). SVMs are very useful in text classification, due to very high dimensional spaces being common in these types of problems. Furthermore, SVMs can be used in either regression or classification problems, despite being more commonly used in the latter (Burges, 1998).

Naïve Bayes classifiers

Naïve Bayes (NB) classifiers work by assuming that the features (which are being used to classify) are conditionally independent, given the class label. This is a very strong assumption that results in classifiers that work well. Despite the assumption not being valid in most of the cases, this model is quite simple, causing it to not suffer from overfitting, making it effective even when operating under an assumption we know is not true (Murphy, 2012).

This technique works well with high dimensional data, since the probability of each feature is estimated independently. In fact, NB classifiers were introduced as a method for text classification in the early 1960s (Maron, 1961).

There exist several algorithms for training NB classifier, with all of them sharing the independence assumption. Some of these algorithms can be trained in ways that exploit closed-form expression evaluation, which is faster than the iterative approaches used by other types of classifiers (Caruana & Niculescu-Mizil, 2006). A closed-form expression is a mathematical expression which can be solved in a finite number of steps.

The NB classifier will make the correct decision as long as the correct class is the most probable one (in the model). The quality of the probability estimate does not matter, as long as it is correct, meaning that the overall classifier can be robust enough to ignore serious deficiencies in its underlying naive probability model (Rish, 2001).

3.2.3 Unsupervised Learning

Artificial Neural Networks

Artificial Neural Networks (ANNs) are an abstraction of a biological system, the neural network of the brain. Despite being much simpler than the brain's neural network, ANNs share two important characteristics with it, namely: the ability to process information in parallel and the ability to learn and generalise from experience (Maimon & Rokach, 2005).

An advantage of ANNs is that they do not make assumptions regarding the latent structure of the data nor about its generative process. The model is instead largely determined by the structure found in the data analysed (Simon, Deo, Selvam, & Babu, 2016). This, however, leads to a sort of ”black-box” operation, in which the model is not easy to understand (it is a collection of weights for the layers), therefore producing little insight about the problem itself. There are several different models of ANNs, as they have been developed over the years by different researchers. The most widely used is the multi-layer feedforward neural network, also called multi-layer perceptron. These types of ANNs are well suited for moddeling relationships between a set of input and output variables (Maimon & Rokach, 2005). A Multilayer Perceptron (MLP) is a class of ANN which uses backpropragation for training. It is a network composed of a number of highly interconnected computing units, called neurons, that are organised in layers. Each neuron processes information by transforming input into processed output. Knowledge is generated and stored in the links of the neurons, under the form of link weights. There is no feedback from the output (Schmidhuber, 2014).

Information processing is done in two steps: first the inputs are converted to a weighted sum of the inputs and the link weights; afterwards, a transfer function is used to convert this sum to an output (Maimon & Rokach, 2005).The purpose of the transfer function is to allow the learning of non-linear relationships.

There are several choices for transfer functions. The most common are: logistic function, hyperbolic tangent and the identity function. The choice of transfer function depends on the type of problem, specifically on the output of the problem (e.g continuous or discrete) (Jordan, 1998).

The link weights are the parameters to be learned. This is done through back propagation (BP). The BP algorithm is performed in two steps. In the forward pass, the predicted outputs given the inputs are evaluated. In the backward pass, partial derivatives of the cost function with respect to the different parameters are propagated back through the network. The network weights can then be adapted using any gradient-based optimisation algorithm.These methods use the gradient of a function at each point in order to guide search. The whole process is iterated until the weights have converged, meaning they stop changing due to reaching a minimum. There is no guarantee of said minimum to be global (Haykin, 1998). This architecture requires lots of input-output examples in order to be trained (Alsmadi, Omar, & Azman, 2009).

The objective function being optimized is typically one that measures the overall error such as mean squared errors or sum of squared errors. The goal of training is to find the set of weights which minimise the objective function (Maimon & Rokach, 2005).

Back propagation algorithms minimise an error function by tuning the modifiable parameters of a fixed architecture, which needs to be set a priori. The MLP performance will be sensitive to this choice: a small network will provide limited learning capabilities, while a large one will induce generalisation loss (Rocha, Cortez, & Neves, 2007),

ANNs are interesting for text analysis because they have the ability to achieve good performance in several tasks without the need to embed explicit knowledge (either semantic or syntactic) (X. Zhang & LeCun, 2015).

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a kind of neural network. They are a regularized version of a MLP. CNNs take advantage of the hierarchy present in patterns in data and use it to recognise and identify more complex patterns. They have an input and output layer and multiple hidden layers, akin to an ANN. There are multiple types of hidden layers, namely: convolution layers, pooling layers, fully connected layers and non-linearity layers (Y. Zhang & Wallace, 2015). While both convolution and fully-connected layers both have parameters that need to be set by the user, pooling and non-linearity layers do not require any.

The convolution layer performs a convolution operation. A convolution is a linear operation that outputs a function that describes how the shape of one input function is modified by another input function. Applying to text multiple inputs can be chosen, for instance, word embeddings can be used as a word representation. In order to represent a five word sentence the input would consist of a $5 \times n$ matrix, where n represents the dimensionality of the word embedding (Y. Zhang & Wallace, 2015). The topic of word embeddings will be approached later in this Literature Review.

Afterwards a non-linear activation function is applied to the output, $tanh$ can be used for this. These layers help the network develop sensitivity for non-linear relationships between input data. These are the non-linearity layers.

Pooling layers perform a pooling operation and are typically applied after the convolution layers. They subsample the output by grouping inputs together. The most used is Max pooling, in which the maximum input value is chosen for a specific pool. Pooling is used as a way to have a standard output size while reducing it, and keeping the most relevant information (Albawi, Mohammed, & Al-Zawi, 2017).

The final layer is the fully-connected one, working similarly to all the layers previously described for ANNs. They are extremely effective while not requiring many pre engineered features, having achieved several state of the are results in natural language processing across several datasets. Along with this, the fact that they use operations that are implemented at hardware levels in GPUs makes them very computationally efficient (Kim, 2014).

For text analysis CNNs architecture is most compatible with classification tasks. These networks will not be applied to this work, as they have been shown to perform much better for longer text. An interested reader is suggested (Y. Zhang & Wallace, 2015) for a good introduction to text analysis using a CNN architecture.

Clustering

Clustering is one of the main tasks of unsupervised learning. The tasks consist of grouping sets of objects such that objects in the same cluster are more similar to each other than to those in other clusters. Similarities in the input data are used to group said data. There are several algorithms to perform this task, such as: k-means, hierarchical and subspace clustering (Murphy, 2012).

K-means Clustering

This method clusters the objects so that each object belongs to the cluster with the nearest mean (which serves as a prototype of the cluster). The mean is the centroid of the objects belonging to the cluster, in this algorithm. The centroid is the average position of the points of a cluster. This problem is computationally intractable, making it so that such problems are usually solved by heuristics. The most popular heuristic involves a simple iterative scheme: first, an algorithm assigns each observation to the cluster whose mean is the ”nearest” (the concept of distance can be represented with different metrics, generating a variation of the algorithm), and then a new mean is calculated (Kanungo et al., 2002).

Proceeding in this fashion will reach a state in which the assignments no longer change, the algorithm has converged. There is no mechanism to ensure that the convergence is optimal (Kanungo et al., 2002).

In order to use the algorithm one must specify k the number of clusters onto which the data will be fit. Also of note is the sensibility of the algorithm to local minima, that is, depending on the initial positions of the initial clusters the results might change dramatically. This can be circumvented by running the algorithm multiple times, while changing both the initial positions and the value of k in order to find the best combination. This process is sadly very computationally expensive (Marsland, 2014).

K-means is an example of a non parametric algorithm, meaning we must choose a fixed k, for each run of the algorithm. In contrast, the following class of algorithms is comprised of parametric ones, in which the number of clusters is inferred from the data.

Hierarchical Clustering

Hierarchical Clustering (HC) seeks to build a hierarchy of clusters. This can be done in an agglomerative or divisive fashion. In agglomerative clustering, the algorithm considers each individual object to be a cluster. Afterwards, these clusters are merged until the desired structure is obtained (the structure is the desired number of different clusters). Divisive clustering works by grouping all objects under the same cluster, which is then iteratively divided until the desired structure is obtained (Maimon & Rokach, 2005). Both HC methods output a dendrogram that represents the nested grouping of objects and the levels at which their similarities change. The merging or division of a cluster is controlled using similarity measures that optimise a criterion. The different configurations between these factors generate various HC algorithms. For more see (Maimon & Rokach, 2005).

Distance metric

The cosine distance is a popular metric for textual content (Sahu & Mohan, 2014).
For two vectors the cosine similarity is calculated as follows:

$$\cos(\mathbf{v1}, \mathbf{v2}) = \frac{\mathbf{v1}\mathbf{v2}}{\|\mathbf{v1}\|\|\mathbf{v2}\|} = \frac{\sum_{i=1}^{n} \mathbf{v1}_i \mathbf{v2}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{v1}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{v2}_i)^2}} \tag{3.1}$$

This procedure operates on a document term matrix (dtm), which is essentially a list of words (each column is a word, each document is a row). By transposing the dtm and applying the similarity formula we get the similarity values between all the words in all the documents. The dtm is weighted using *tf-idf*. This measure is introduced in section 3.2.7.
It should be noted that this is a similarity measure, in order to use it for the K-means algorithm the inverse is taken first, leading to:

$$dist = 1 - \cos(\mathbf{v1}, \mathbf{v2}) \tag{3.2}$$

3.2.4 Natural Language Processing

Natural Language Processing (NLP) is a field of research characterised by the investigation and development of a set of methods that make natural language accessible to computers. Its focus is the design and analysis of algorithms and representations for processing natural language (King & Reinold, 2008).

As a field of research, NLP is closely related to the fields of Artificial Intelligence (AI, the broad field of making a machine intelligent, that is, capable of making human-level decisions (Russell & Norvig, 2009)), and machine learning. All these fields are related between themselves and, more broadly, they are related to the field of computer science (King & Reinold, 2008).

The ultimate goal of NLP is an ambitious one, namely the conversion of text into a programmer-friendly data structure that describes the meaning of natural text (Ronan et al., 2011). Until this fundamental problem is solved, we must settle for the reduced objective of extracting simpler representations that describe limited aspects of the textual information. Since unstructured text remains the largest readily available source of information, NLP research has been on the rise in recent years (Gupta & Lehal, 2009).

Natural language processing is comprised of a set of well-defined tasks that are currently being used to solve various problems. Examples are Sentiment Analysis, Topic modelling, Named Entity Recognition, etc (Nitin & Fred, 2010).

The work at hand deals with analysing surveys which have open-ended questions and the possibility for the respondent to leave comments. Essentially, this work sets out a proposed methodology for the analysis of open-ended questions in questionnaires, proposing the use of automated tools to perform the analysis. Open-ended questions, despite harder to analyse, are considered to provide a more accurate view of the respondents own thinking (Geer, 1991).

In the next subsections we will introduce fundamental NLP concepts as well as more advanced techniques that have potential applications for solving the problems addressed in this work.

3.2.5 Fundamental concepts of NLP

In processing text with the aid of computational methods, assumptions will inevitably be made. In this section we will mention the more general ones, mentioning some properties of words in a document, namely their distribution, which obeys a power law (King & Reinold, 2008).

Bag-of-words model

The Bag-of-words (BOW) is a simplified way of representing text. It records two aspects of the original text, namely the words that occur and their frequency, with the ordering of the words being ignored (Ljungberg, 2017). This model, despite not realistic, since in a document the order of words is important, is extremely useful, as it allows text to be represented in a way that is easy for computers to manipulate. This results in a very sparse vector, due to the distribution of words in text. Firstly, studying the statistical patterns of human word usage can be used to understand meaning. The study of this phenomena is called statistical linguistics (Turney & Pantel, 2010). This is the encompassing concept for the distributional hypothesis stated below.

Zipf's law

Zipf's law is named after George Kingsley Zipf, an American linguist. He made it a popular concept, and sought to explain it, despite not claiming to have originated the idea (Powers, 1998).

Applied to text, the law states that the frequency of a word $f(w)$, is a nonlinearly decreasing function of the rank of the word $r(w)$, in a corpus. In this context, rank refers to the position of a specific word, in a decreasingly ordered list of frequency (number of occurrences in the corpus). The corpus is the entire collection of texts being analysed (Kumar, 2016).

$$f(w) = \frac{C}{r(w)^a} \tag{3.3}$$

C is a constant determined by the particulars of the corpus, the frequency of the most frequent word.

Zipf's law is important because it provides a baseline model for the expected occurrence of target terms. This law provides a distributional foundation for language models, and permits their evaluation (Brent, 1997).

Distributional Hypothesis

The distributional hypothesis states that words occurring in similar contexts tend to have similar meanings. This hypothesis is the justification for the use of word embeddings (Zellig, 1954), which will be discussed later in the Literature Review.

The distributional hypothesis can be extended: instead of applying it to words, it can be applied to dependency trees of a parsed corpus. The extended distributional hypothesis states that patterns that co-occur with similar pairs, (X,Y) in this case, tend to have similar meanings. It was devised as a way of generating inference rules, in an unsupervised fashion. An inference rule is a rule for making inferences from textual data, such as *'X wrote Y'* $\approx$ *'X is the author of Y'*. (Mcdonald & Ramscar, 2001)

Latent Relation Hypothesis

This relation is the inverse of the extended distributional hypothesis. It states that pairs of words that co-occur in similar patterns tend to have similar semantic relations. For example *mason:stone, carpenter:wood,potter:clay* share the semantic relation *artisan:material* (Peter, 2008).

3.2.6 Text Pre-Processing

Extraction

In order for the text to be read by a machine, several steps are necessary. The goal of these steps is to convert the strings of text into data frames composed by the individual word counts. This is a classic format for numerical data. As such, analyses, visualisation and management of the data is made easier and consistent with already existing tools for data analysis.(Silge, 2017) The tables take the form of one token per row.

Tokenization

In order to convert text into tokens, tokenization is performed. Tokens can have an arbitrary length and are used in the downstream NLP tasks, such as building a language model. In fact, changing the length of the tokens creates a new type of model. When one word corresponds to one token we use a unigram-based language model, while in the case of two words corresponding to a token we have a bigram based language model (Silge, 2017). An option is to include punctuation and other special characters. These have different uses and as such influence the choice (Kumar, 2016).

The sparsity of the resulting matrix, that is the number of rows that are 0, is very large. Most words do not appear in most documents of a large collection. This also has to do with the 'burstiness' phenomenon (Doyle & Elkan, 2009). Both these factors contribute to having sparse matrixes as input.

To improve computational efficiency these rows can be suppressed without losing much information. This step is performed only for increasing the computational efficiency of the following calculations (Silge, 2017).

Normalisation

After collecting the tokens they must be normalized. This is normally done by "stemming" or "lemmatization" and converting all words to lowercase. Stemming is the process of cutting the end of words, normally derivation affixes, such as the "-*s*" commonly added to form the plural of a word. Lemmatization is a more sophisticated form of stemming, in which words are reduced to their base form, through the use of dictionaries, for example. This is done to facilitate the analysis of data. (Kumar, 2016)

Finally, in order to have a cleaner input for the following tasks, stopword removal is performed. Stopwords are functional words that do not convey new information or meaning to a sentence, such as *"the", "a", "and"* and many more. An easy approach is the use of a dictionary containing a list of stopwords, and then removing them from our matrix. There are several lists for the English language and according to the field of study, some specific words, are easily added (Schofield, Magnusson, & Mimno, 2017).

During normalisation procedures, care must be taken not to prune the data too much, in order to limit the loss of information (Schofield, Magnusson, Thompson, & Mimno, 2017).

3.2.7 Standard NLP activities

In this section, we shall expose the techniques which will be applied. They were chosen for their applicability to an analysis of data with an exploratory nature. The presented techniques allow for hidden patterns to emerge from the textual data available, making them suitable for the purpose of the case study, which will be presented in the Research Methodology section.

Descriptive corpus statistics

A common first step of a text analysis pipeline is the extraction of statistical information regarding the corpus. The corpus is the collection of documents, where a document corresponds to an individual unit of text, such as a journal article or a questionnaire, to be analysed.

One such metric is word frequency, which is simply a count of the number of times a specific word appears. This information can intuitively point towards relevant concepts or topics for a document (informative words used more often tend to be more important).(Silge, 2017)

The heuristic 'term frequency'-'inverse document frequency', *tf-idf*, is a measure of how important a word is to a document in a collection (or corpus) of documents. *tf-idf* consists of two separate statistics that are multiplied together, and in which each term captures a different concept about the statistical distribution of text (Jurafsky & Martin, 2000).

The tf is calculated as

$$tf(W) = n/V \tag{3.4}$$

where n is the number of times word W appears, and V is the total number of words across all documents. The division by V is a way of normalizing the frequency for larger documents.

The idf is calculated as

$$idf(W) = ln \frac{n_{documents}}{n_{documents\ containing\ w}} \tag{3.5}$$

This equation gives us a measure of how much documents the word is in, allowing us to weigh down terms that appear in most documents, as these are less informative (Nguyen, 2014). This is a widely used weighting scheme in information retrieval (Silge, 2017).

Visualisation with Word Clouds

Another type of analysis that can be performed recurring to just simple corpus statistics is the "word cloud".

Word clouds are a straightforward and visually appealing method for visualising text. They provide an overview of the text by depicting the most important words in it. The measure of importance can be different. Frequency can be used, but so can *tf-idf* and other weighting schemes. (Heimerl, Lohmann, Lange, & Ertl, 2014)

Word clouds are useful because they can serve as a starting point for analysis. However, they do have the drawback of not taking linguistic knowledge into account. Word clouds are useful to give an impression of what information is present on the text (Kuo, Hentrich, Good, & Wilkinson, 2007).

There are several developed systems which make use of word clouds as a static way to visually summarise documents (Wu et al., 2010) (Stasko, Gorg, Liu, & Singhal, 2007).

Sentiment Analysis

The main goal of sentiment analysis is to assign a score to a sentence. This is called the sentiment score. For instance a sentence can be scored as *positive, negative, neutral.* This analysis allows us to interpret reviews of products, surveys and even news articles, in order to quickly discover what the general feeling regarding our subject of focus is, which is useful in the context of analysing surveys (L. Zhang & Liu, 2016). Besides identifying the opinion itself, these types of techniques also usually identify the subject, the topic that is being discussed, and the opinion holder, which is the person that holds the opinion. Furthermore, the system must be able to differentiate between objective and subjective sentences, which is a subtask in sentiment analysis.

There are several configurations to perform such analyses, namely, the first parameter impacting the process being the choice of scope. Sentiment analysis can be performed at document, sentence, and sub-sentence levels, which gives us a different level of detail. The detail increases from document to sub-sentence levels as the latter can express more opinions in a single document (Devika, Sunitha, & Ganesh, 2016).

From a more technical view, the different approaches can be grouped into four classes, namely: machine learning, lexicon-based, statistical and rule based approaches. Machine learning is the most commonly used method. However it requires a significant data set for training, which may be unavailable (Anaıs, Crina, Damien, Omar, & Lionel, 2013). In Collomb's work one can also see that a rule-based approach can be highly effective.

Lexicon-based approaches use NLP and lexical resources to assign sentiment. They use mainly POS information and WordNet (Anaıs et al., 2013). These approaches create a sentiment lexicon in which words are scored according to their sentiment. Afterwards, the words in text are compared to this lexicon (Nielsen, 2011).

Using a simple lexicon-based approach tends to produce poor results. This happens since there are sentences which express strong opinions without using those intuitive words contained in the lexicon (Dalal & Zaveri, 2014). Another problem with lexicon-based approaches arises when a lexicon created for a specific source of document is applied to another source. For instance, a lexicon developed for novels will not perform well when dealing with financial news (Loughran & McDonald, 2011b).

Topic Models

Topic models are statistical models that learn the latent structure in document collections. The original generative model is Latent Dirichlet Allocation (LDA). The LDA model assumes that documents have multiple topics. In LDA each of the D documents is modelled as a discrete distribution over T latent topics. Each topic is a discrete distribution over the vocabulary of words W. The number of topics to fit to the data must be fixed before running the algorithm in traditional LDA (Airoldi, 2014). From this original model, topic models arose as a class of probabilistic models that has become a central subject of research in text mining, computer vision and bioinformatics (Heinrich, 2009).

In order to select the best performing number of topics for the corpus being analysed, several metrics exist. Such metrics are intensive computationally, as they require multiple LDA models to be trained in each iteration (Arun, Suresh, Madhavan, & Murthy, 2010).

A typical assumption made when applying topic modelling is that the latent space is semantically meaningful. Empirically, it has been verified that topic models do lead to good models of the documents. Topics tend to place a high probability on words representing concepts and documents tend to be represented as an expression of the previous concepts (Chang, Boyd-graber, Gerrish, Wang, & Blei, 2009).

The generative process for the text could be the following: first a distribution is randomly chosen among topics, then, for each word in the document, randomly choose a topic from the distribution over topics and randomly choose a word from the corresponding topic. One should note that words are generated independently from other words, which is a simplifying assumption (D. Blei, 2006).

LDA is a mixed membership model, meaning that each document can have more than one topic assigned to it. This representation is more flexible than models where only one topic is assigned to a document, but leads to a much tougher optimization problem during model training (one with multiple optima) (E., M., & Dustin, 2016). Having more than one local optimal value means that LDA is sensitive to starting parameters.

Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a statistical method, based on corpus-wide statistics, for inducing and representing certain aspects of words and passages, reflected in how they are used (L, Laham, Rehder, & Schreiner, 1999).

The idea behind this method is that the aggregate of all word contexts, wether a word appears or not, largely determines the similarity of meaning among words or sets of words. LSA produces measures of word-word, word-passage and passage-passage relations, well correlated with human judgments. It should be noted, however, that these results depend on the chosen dimensionality for the representation, which must be heuristically set (Landauer, Foltz, & Laham, 1998).

LSA works by applying singular value decomposition (SVD) on a matrix, in which each row stands for a unique word and each column contains context. Context can be another word, a passage, or even a sentence. This controls how fine grained the analysis is. In this work, word-word co-occurrence matrixes are used. A weighting scheme (for instance, *tf-idf*) is applied to the matrix before performing SVD. These results are then truncated, with this step being the one where the dimensions are reduced.

The dimensionality reduction step is a form of induction, allowing for added information to be extracted from mutual constraints among a large numbers of words in a large number of contexts. The words are then represented as vector components, with normally 100-500 dimensions (L et al., 1999).

LSA represents passages by summing up the word vectors contained in the passage, without regarding word order. This information can then be used to compare different answers, measuring how similar they are to one another. This is done using a distance measure between the vectors we wish to compare, usually the cosine distance is the used distance measure (L et al., 1999).

LSA is interesting regarding case study objectives, as it has been previously applied towards the automatic analysis of open ended questionnaire responses ((Martin, Martin, & Berry, 2016); T. Leleu (2008) ;T. D. Leleu et al. (2011)).

Word embeddings

Word embeddings are vector representations of text. They project the high-dimensional sparse word occurrence data onto denser and lower dimensional vectors, the embeddings (Almeida & Xexéo, 2019).

Word embeddings are important because they encode accurate syntactic and semantic word relationships and they can be used as features in other NLP tasks (Mikolov, tau Yih, & Zweig, 2013). Many of these relationships can be represented as linear translations of the word embeddings (Tomas, Ilya, Kai, Greg, & Jeffrey, 2013)

According to the method used to generate word embeddings, they can be classified either as prediction-based or count-based embeddings (Almeida & Xexéo, 2019).

Prediction-based models leverage information about the most probable next word (much like language models), as a way of assigning features to each word in the vector space (Tomas et al., 2013). Count-based methods, on the other hand, leverage global statistics about the corpus (such as word co-occurrence) for feature assignment (Pennington, Socher, & Manning, 2014).

Word embeddings are good at disambiguation tasks, since they are very good at synonym detection (synonyms tend to be very close together on the projected vector space) (Baroni, Dinu, & Kruszewski, 2014). They are also good for named entity recgonition, the extra features generated increase performance through a better representation (Almeida & Xexéo, 2019).

3.2.8 Pre-processing and *tf-idf* example

In this section, a representative example of these operations is given, in order to facilitate the understanding of how the data structures used work and interact with each other.

For the example, sentences we can use are '*The quick brown fox jumped over the lazy dog.*' and '*This red dog jumped over the lazy fox*'. These sentences are enough to demonstrate how these operations take place.

In this example we are using a unigram model, meaning that each word corresponds directly to one token. The token for '*quick*' is simply *quick*. For a bigram model, the tokens would be: *The quick* , *quick brown*, and so on until all two word sequences are represented in the matrix. Using a bigram model creates an even sparser matrix, because two word sequences are less frequent than 1 word occurrences (Tan, Wang, & Lee, 2002). See tables 3.1 and 3.2.

The bag of words model consists of representing all tokens present, the vocabulary, as well as their frequency counts in different documents. The two sentences provided are taken to be the documents for this example, with both of them together constituting the corpus.

The bag of words representations for these simple sentences is shown in the following table. Together with the bag of words representation, a *tf-idf* weighting is applied, in order to demonstrate the metric.

Unigrams	Sentence 1	Sentence 2	*tf-idf*
red	0	1	0,06931472
this	0	1	0,06931472
the	2	1	0
quick	1	0	0,06931472
brown	1	0	0,06931472
fox	1	1	0
jumped	1	1	0
over	1	1	0
lazy	1	1	0
dog	1	1	0

Table 3.1: Bag of words representation of the example sentences. S=0.20

In the following table a bag of bigrams is represented. As we can see, from the same two sentences a sparser matrix is generated (more zeroes in the rows of the columns corresponding to bigram frequency, columns two and three).

Bigrams	Sentence 1	Sentence 2	*tf-idf*
the quick	1	0	0,06301338
quick brown	1	0	0,06301338
brown fox	1	0	0,06301338
fox jumped	1	0	0,06301338
jumped over	1	1	0
over the	1	1	0
the lazy	1	1	0
lazy dog	1	0	0,06301338
the red	0	1	0,06301338
red dog	0	1	0,06301338
lazy fox	0	1	0,06301338

Table 3.2: Bag of bigrams representation, with *tf-idf* scaling applied. S = 0.37

Sparsity is calculated as the number of zero elements, divided by the number of total elements of a matrix (the *tf-idf* column does not count towards this result). The S value in the table captions is the sparsity of each bag of tokens representation. Sparsity is an important concept, because it allows for the creation of more efficient computer algorithms and data storage structures, by avoiding the operations involving the zero entries in the matrix (Gilbert, Moler, & Schreiber, 1992).

As an example of stemming, the word '*jumped*', when stemmed corresponds to only its root, '*jump*'.

3.3 Summarising remarks

Taking into account the literature review, we can state that there exist a multitude of techniques from various research areas, such as NLP and AI, that can be applied in order to fulfil the goal of the current work.

In this literature review it is possible to see the variety of methods available to perform any single task. Accounting for previously identified challenges, in the Introduction and Problem Definition, and with the goal of the analysis in mind, we can select the most appropriate techniques to perform the analysis, which will be done in the research methodology.

The present chapter serves as a backbone for the analysis process. It allows for the informed and justifiable choice of methods, while making evident some challenges that might arise during the analysis. These challenges are discussed in detail in the closing remarks.

Chapter 4

Research Methodology

In this chapter, in order to meet the overall objective of the work, the most appropriate methods and techniques from the Literature Review are selected, making explicit the design of the tool. Furthermore, the data acquisition process is explained and the main outputs of data analysis listed and explained.

4.1 Overview

The proposed methodology consists of various steps that in the end enable the efficient development of personas. First and foremost the data was be collected, organised and cleaned.

In order to collect the data an online survey was be developed using the Google Forms platform. The questionnaire was handed out with the help of IST student's organisations via e-mail and posted in Social Media platforms. Answers were collected during a multi-week period, from 22/04/2020 up to 09/07/2020.

Once all the data is correctly read and stored the tool is ready to be applied. In figure 4.1 we can see a flowchart for this process. Stages can be associated with chapters and sections of this document. Stages two through four are heavily related with this chapter, in fact, the ordering of the figure matches the ordering of subsections. Stage two corresponds to section 4.2, stage three corresponds to section 4.3 while stage four corresponds to sections 5.1 and 5.2.

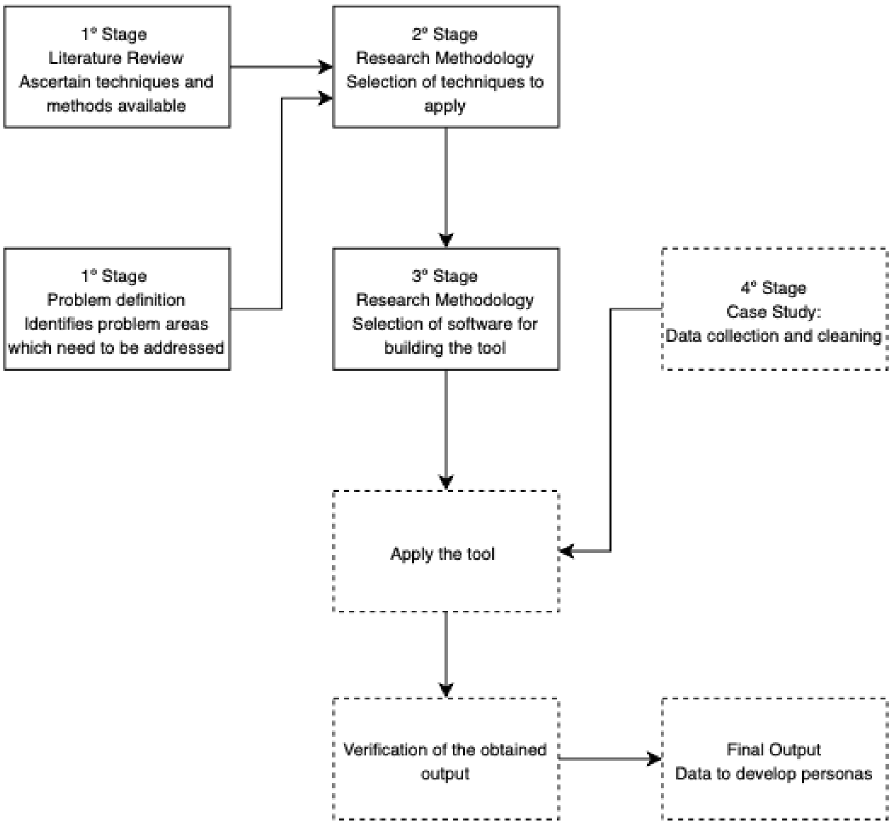

1° Stage
Literature Review
Ascertain techniques and methods available
2° Stage
Research Methodology
Selection of techniques to apply
1° Stage
Problem definition
Identifies problem areas which need to be addressed
3° Stage
Research Methodology
Selection of software for building the tool
4° Stage
Case Study:
Data collection and cleaning
Apply the tool
Verification of the obtained output
Final Output
Data to develop personas

Figure 4.1: Roadmap for the design of the tool

This figure shows how the research work here presented fits into the technical side of this work, relating the research activities with the technical steps of designing and applying the tool to a specific dataset.

4.2 Tool workflow

The proposed tool will have a sequential workflow until a normalized representation of the text is obtained.

As we can see in the figure below, the first steps of the analysis are concerned with extracting and transforming the important data, using procedures discussed in the Literature Review, into a format that is compatible with the chosen computer software.

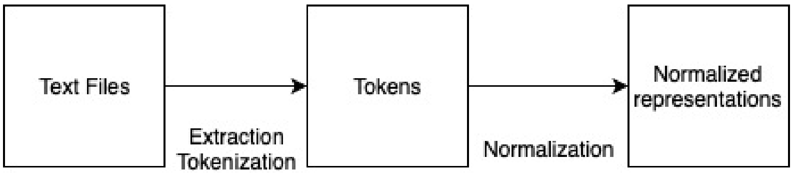


Figure 4.2: Overview of the the data collecting and cleaning workflow

The normalized representations are the main output of the data collection stage. They are the starting point for performing data analysis. In order to obtain these representations from the collected text files, the following procedure is followed. While the files are being read, tokenisation is performed, leading to a list of token counts being obtained. The tokens are then lowercased. Stopword removal is done only for specific tasks, and as such is performed right before these, while preserving the original data with stopwords.

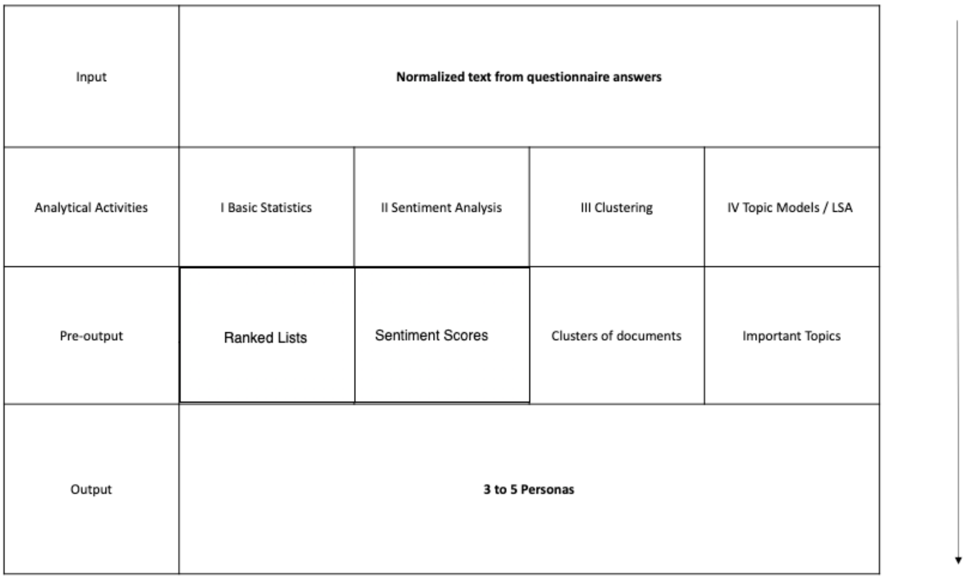
Taking the normalized representations generated as input and using techniques from NLP, the analysis is performed. In the following table we can see which pre-processing steps were applied to each analytical task performed.

Pre Processing / Task	Simple Statistics	Sentiment Analysis	Topic Modelling / LSA	Clustering
Unigram Tokens	X	X	X	X
Bigram Tokens		X		
Remove Stopwords	X	X		
Reduce Sparsity	X	X	X	X
Lowercasing	X	X	X	X
Stemming				

Table 4.1: Pre-processing done per analysis task

In figure 4.3 the output of each task is shown. Different techniques and software packages were used to produce visualisations of the different results.

The following figure shows how the analysis proceeds. It goes from top to bottom, and the analytical activities are performed from left to right. Starting with simpler analysis grants insights about the data that are useful for the analysis, this is the only reason why these are performed in the first place.



Input	**Normalized text from questionnaire answers**			
Analytical Activities	I Basic Statistics	II Sentiment Analysis	III Clustering	IV Topic Models / LSA
Pre-output	Ranked Lists	Sentiment Scores	Clusters of documents	Important Topics
Output	**3 to 5 Personas**			

Figure 4.3: Overview of the text analysis workflow.

In figure 4.3 we can see that three main techniques were selected, these being Sentiment Analysis, Clustering and Topic Modelling. These were chosen based on 2 characteristics: they are unsupervised, meaning that these techniques explore the data without needing an annotated dataset (the base case). The second reason is that these techniques are closely related with the goal of creating personas, this is, they uncover hidden similarities among the data.

Following the data analysis there was a persona development step, in which personas were created according to the analysis. These personas can then be used to communicate the findings effectively, as well as focus on the goals and needs of the characterised population (Cooper, 1999).

4.3 Software packages selected

This methodology proposal is designed to be able to perform a fast and efficient analysis of answers to short open-ended question surveys. The analysis tool was based on R language scripts, using packages that are specifically designed for text analysis. R is an open-source programming language designed for statistical analysis (Welbers, Atteveldt, & Benoit, 2017).

In table 4.2, the packages used for each task are presented, as well as the CRAN link of each package. CRAN stands for The Comprehensive R Archive Network, which is a network of servers where R packages are stored and maintained. These packages are accompanied by vignettes that explain their functions, and, when available research papers explaining the concepts used. All this is available in the links provided in table 4.2.

	Tasks	Packages	CRAN link
	Extraction	base R	
Pre-processing	Tokenization	tm	https://CRAN.R-project.org/package=tm
	Normalization	tm	https://CRAN.R-project.org/package=tm
I	Basic Statistics	tidytext	https://CRAN.R-project.org/package=tidytext
II	Sentiment Analysis	SentimentAnalysis	https://CRAN.R-project.org/package=SentimentAnalysis
III	Clustering	cluster	https://CRAN.R-project.org/package=cluster
IV	Topic modelling	topicmodels	https://CRAN.R-project.org/package=topicmodels

Table 4.2: Packages used to perform analytical activities.

There are several packages that perform the same tasks. The ones presented here were chosen as they integrate well with each other, using the same data structures for input/output therefore easing the task of data analysis.

4.4 Persona generation

Following the application of the tool to the dataset, a grouping of the answers was obtained. From this grouping personas were developed. The data was analysed in an exploratory fashion, several clusterings were applied. Documents were clustered based on word similarity alone to exploit the distributional hypothesis.

Sentiment scores were computed for each answer. From this a table of sentiment scores was created and analysed.

Demographic statistics were computed for each group of respondents. This helped in choosing the fictional student to represent each persona.

Finally a narrative was constructed for each persona. This is the least automatised step of the process.

Chapter 5

Results and Persona Design

5.1 Case Study

The proposed work consists of a methodological development, a case study which allowed the testing of said methodology is required. In order to do so, and in-line with the recent report by Comissão Análise do Modelo de Ensino e Práticas Pedagógicas, a questionnaire was developed that elicits information about how the students feel about the current teaching practices and how prepared they think they are for the future as professionals.

Summarising, this case study allows both for the methodology to be tested, while at the same time allowing for the collection of relevant information.

5.2 Data and questionnaire

The data to use in this work consisted of answers to an open-ended questionnaire. Methodological issues regarding this type of data have been discussed in both the problem definition and the literature review.

Since the data contains personal information, a clearance by the Ethics Commission of Técnico was obtained to perform this work, here in Appendix C.

A questionnaire was used as the data collection tool. The questionnaire consisted of five questions that allowed us to extract information relevant towards characterising the students. Here we will present the questions as well as the reasoning behind why a question is posed.

The overall reasoning for this type of questions comes from the fact that they elicit subjective information from the respondent, such as sentiments about a specific topic (Geer, 1991). This information is crucial in defining a persona, as the goals and objectives of each persona created must be defined (Cooper, 1999).

By asking respondents how they feel or to perform comparisons among the respondent's peers, we hope that the answers are subjective and sentiment-heavy, allowing the chosen algorithms to perform well and deliver good analytical information.

The first question tries to elicit the respondents to talk about their objectives in attending college. This is a good differentiator between students, as each one will have a somewhat different motive for attending college. This question is posed as: ”*I go to university because...*”

In the second question, which is posed as ”*When thinking about Técnico I feel...*” information that is a strong differentiator between students is elicited. This question, in a very simplified way, will tell us how the students feel about Técnico.

In the third question the respondent is asked about the skills gained during its education. This question serves both to elicit expectations and beliefs about the education provided. This question is posed as ”*The skills I acquire at Técnico enable me to...*”

The fourth question asks about the future workplace of the student. This question tries to assess how the student is feeling about the education being provided. In this question, information regarding both the present and the future is collected. It serves to assess the expectations of the students. This question is posed as ”*In my future workplace I want to...*”

The fifth question is meant to help in understanding how students feel about the current world situation. This question is posed as ”*When thinking about the current COVID-19 situation, I feel...*”

The final question is also meant to help in in understanding how students feel about the current world situation. This question is posed as ”*Living the current COVID-19, made me want to...*”

This information is summarised in table 5.1, along with it, the temporal loading of each question is shown as well as the behavioural characteristic that we try to capture with each question, such as a belief or motivation of each individual respondent.

Question	Prompt	Behavioural Characteristic	Temporal Loading
I	I go to university because...	Motivation	History/present
II	When thinking about Técnico I feel...	Needs / Frustrations	Present
III	The skills I acquire at Técnico enable me to...	Motivation / Goals	Future
IV	In my future workplace I want to...	Goals	Future
V	When thinking about the current COVID-19 situation, I feel...	Frustrations	Present
VI	Living the current COVID-19, made me want to...	Frustrations	Present

Table 5.1: Questionnaire

Dataset Characteristics

In this work, 162 questionnaires were collected. The questionnaires were made available on the Google Forms platform, and they were sent to respondents through e-mail and social media platforms. The data collection period was from May to July 2020. The questionnaires were disseminated via e-mail and social network posts. The e-mails were sent to students of each degree by the corresponding student organisations. The social network posts were posted and subsequently re posted, as a reminder, in a student group of IST, by the author. Of these, 31 were not suitable for analysis due to various factors, leaving us with 131 to be analysed.

From IST's[1] website there are 10.468 students enrolled, meaning that responses from approximately 1,25% of the student population were collected.

This number, despite not being sufficient to give us a very clear picture about the universe of IST's students is, in hindsight, sufficient to serve as a preliminary test for this methodology. The dataset had, after removal of the answers that were not eligible, 3322 words. Of these words 1102 were removed, as they were stop words. This is the initial stage of the pre-processing necessary to analyse textual data, as mention in the Literature Review section.

The demographics collected regarding the respondents are presented here in the following tables and are the following:gender, year in college and degree which they are attending currently.

Firstly, the distribution of the curricular years is reported in table 5.2.

Curricular Year	Respondents	%
1	16	12,2
2	16	12,2
3	19	14,5
4	28	21,4
5	52	39,7
Total	131	100

Table 5.2: Distribution of the curricular year of each respondent.

In the following table the distribution of respondents by gender and degree is presented.

[1] https://tecnico.ulisboa.pt/en/about-tecnico/institutional/presentation/ consulted 28th august 2020.

Degree	Total (%)	Male	Female	Prefer not to say
Aerospace Engineering (MEAer)	9 (7,1)	6	3	0
Architecture (MA)	2 (1,6)	1	1	0
Biological Engineering (MEBiol)	12 (9,5)	2	9	1
Biomedical Engineering (MEBiom)	12 (9,5)	4	8	0
Chemical Engineering (MEQ)	4 (3,2)	2	2	0
Civil Engineering (MEC)	2 (1,6)	1	1	0
Computer Science and Engineering (LEIC)	9 (7,1)	5	4	0
Electrical and Computer Engineering (MEEC)	30 (23,8)	25	5	0
Engineering Physhics (MEFT)	8 (6,3)	4	3	1
Environmental Engineering (MEAmbi)	4 (3,2)	2	2	0
Geological and Mining Engineering (LEGM)	1 (0,8)	1	0	0
Masters in Industrial Engineering and Management (MEGI)	5 (4,0)	5	0	0
Materials Engineering (MEM)	8 (6,3)	6	2	0
Mechanical Engineering (MEMec)	18 (14,3)	10	6	2
Naval Architecture and Ocean Engineering (LENO)	1 (0,8)	0	1	0
Not Listed	6 (4,8)	4	2	0
Total (%)	131	78 (59,5)	49 (37,4)	4 (3,1)

Table 5.3: Degree and gender distribution.

From tables 5.2 and 5.3 it is clear that the sample is not representative of the actual population under study. This poses as a major challenge towards validating these results. These concerns will be discussed in depth in chapter 6, in a section dedicated towards limitations with the present work.

5.3 Data Presentation

In this section the collected data is presented and some brief analysis is provided. In the following section, all the data will be amalgamated and analysed in a way that allows the building of personas in a subsequent step.

There is no particular order in which the techniques must be applied for this methodology, however it is beneficial to start with a higher level analysis, for instance, ranking words based on a weighting scheme, as this can quickly point towards the right direction in subsequent analysis. As such the results were presented and discussed in the same order that they were introduced in the text in the Research Methodology section. Furthermore, the cluster assignments and the output from the LDA parameter optimization are presented respectively in annexes A and B.

Basic Statistics

After this preliminary characterisation the analysis starts.
Firstly, ranked lists are created containing the most informative words, according to the *tf-idf* metric.

These are presented in the following tables.

The *tf-idf* value being the same for different words implies equal count statistics for both words (number of times it appears in a document and number of documents is appears in).

This preliminary analysis is a starting point. In the following analysis these points will be further explored and discussed.

For the ranked lists base functions from R and from the package *'tidytext'* are used.

Term	*tf-idf*
supposed	5,26
graduate	5,26
important	5,26
stuff	4,85
ambitions	4,85
keep	4,85
degrees	4,85
bright	4,85
urge	4,85
engineer	4,85

Table 5.4: List of highest ranking words present in answers to question 1, by *tf-idf*.

Term	*tf-idf*
sadness	4,70
motivated	4,70
stress	4,42
revolted	4,41
lost	4,41
smart	4,35
contempt	4,35
frustrated	4,35
slightly	4,32
education	4,25

Table 5.5: List of highest ranking words present in answers to question 2, by *tf-idf*.

Term	*tf-idf*
resilient	6,06
grow	5,51
ambitions	5,37
research	5,37
trust	4,45
expertise	4,45
difficult	4,11
hopefully	4,11
perform	4,05
autonomously	3,98

Table 5.6: List of highest ranking words present in answers to question 3, by *tf-idf*.

Term	*tf-idf*
thrive	5,81
recognized	5,81
succeed	5,81
rich	5,81
succeed	5,81
dominate	5,81
teach	5,81
autonomy	5,81
confortable	5,70
happy	5,12

Table 5.7: List of highest ranking words present in answers to question 4, by *tf-idf*.

Term	*tf-idf*
smart	5,97
bored	5,97
confident	5,97
tired	5,97
traped	5,97
relaxed	5,97
restrained	5,97
insecure	5,97
anguish	5,97
hate	5,97

Table 5.8: List of highest ranking words present in answers to question 5, by *tf-idf*.

Term	*tf-idf*
exercise	5.87
run	5,87
give	5,60
living	5,18
self	5,18
money	5,18
win	5,18
transparency	4,77
interactions	4,77
outside	4,63

Table 5.9: List of highest ranking words present in answers to question 6, by *tf-idf*.

Sentiment Analysis

Sentiment scores were calculated using the function *'analyzeSentiment'* from R's package *'SentimentAnalysis'*. The scores were calculated using the 'GI' dictionary[2], as it is suited for this type of textual content. More information can be found on the package repository, mentioned in table 4.2.

The scores were calculated individually for each answer. This information allows inference of important differences in personality, for instance, optimistic/pessimistic, which is key for the objective of developing a persona.

To begin our analysis of sentiment the fraction of each score is presented for each answer in the following table. This information allows us to gauge whether the reaction to the question by a specific respondent was in tune with the others or not, again hinting at important information for the development of personas.

Question	1	2	3	4	5	6
Positive	0,4885	0,3053	0,4198	0,4046	0,1756	0,3511
Neutral	0,4736	0,4598	0,4892	0,5727	0,5214	0,5276
Negative	0,0379	0,2348	0,0909	0,0227	0,3030	0,1212

Table 5.10: Fraction of (positive,negative,neutral) sentiment scores sorted by question.

The sentiment scores are presented next as a table. A score was calculated for each individual

[2] GI stands for general inquirer, more information on the dictionary can be found here `http://www.wjh.harvard.edu/~inquirer/`

answer. The scores are such that: $s \in [-1, 1]$ according to how negatively or how positively an answer is rated.

Summing the scores across each row we get an idea of how negative or positive a respondent was towards the questions posed. A higher number means that overall the respondent was more positive. Summing across each columns provides a glimpse into how each question was reacted to by the population of respondents. A higher number indicates that the question elicited a more positive response.

Respondent / Question	1	2	3	4	5	6	Sum
1	0,5000	1,0000	0,3333	1,0000	1,0000	0,2857	4,1190
2	0,2000	0,0476	0,0000	0,2000	-0,1053	0,0000	0,3424
3	0,0000	-1,0000	0,0000	0,0000	0,0000	-0,5000	-1,5000
4	-0,2500	0,0000	0,5000	0,2500	0,0000	0,0000	0,5000
5	0,0000	0,0000	-0,3333	0,0000	-1,0000	0,0000	-1,3333
6	0,0286	0,1250	0,0000	0,1667	0,2000	0,4286	0,9488
7	-0,5000	0,0000	-0,5000	0,0000	-0,2500	0,0000	-1,2500
8	0,1111	1,0000	0,0000	0,0000	-0,1818	0,0000	0,9293
9	0,0000	0,1000	0,0000	0,1429	0,1429	0,1250	0,5107
10	0,0000	0,0909	0,0000	0,1818	-0,0556	0,0000	0,2172
11	0,0000	0,0000	0,0000	0,0000	0,0000	-0,3333	-0,3333
12	0,0000	0,0000	0,2500	0,0000	0,0000	0,0000	0,2500
13	0,0000	-0,1667	0,0000	0,0000	0,0000	0,0000	-0,1667
14	0,1667	0,5000	0,3333	0,0000	0,0000	0,0000	1,0000
15	0,0000	-1,0000	-0,2500	0,3333	-1,0000	0,2500	-1,6667
16	0,0741	0,2000	0,0606	-0,0833	0,0000	0,0000	0,2513
17	0,0000	0,0000	0,3000	0,0000	0,0000	1,0000	1,3000
18	0,0000	0,0000	0,5000	1,0000	0,1111	0,0000	1,6111
19	0,0000	-1,0000	0,3333	0,5000	1,0000	0,2000	1,0333
20	0,5000	0,2500	0,2500	0,0000	0,1250	0,0000	1,1250
21	0,2500	-0,3333	0,3333	0,0000	1,0000	0,0000	1,2500
22	0,0000	-1,0000	0,0000	0,3333	0,0000	0,1667	-0,5000
23	0,0000	0,0000	0,0000	1,0000	0,0000	0,0556	1,0556
24	0,0769	0,0000	0,0000	0,0833	0,0698	0,6667	0,8967
25	0,0000	0,0000	0,0000	0,0000	0,0000	0,1250	0,1250
26	0,0000	-0,0476	0,1667	0,0000	-0,1429	1,0000	0,9762
27	0,0769	-1,0000	0,0000	0,0000	0,1111	0,2500	-0,5620
28	0,0000	0,0769	0,0000	0,1667	-0,3333	0,0625	-0,0272
29	0,0000	0,0000	0,2000	0,3000	0,0435	0,0000	0,5435
30	0,2143	0,0345	0,0909	0,0208	0,0000	-0,0833	0,2772

31	0,1538	0,5000	0,0000	0,0000	0,0000	0,0000	0,6538
32	0,0000	0,0000	0,1111	0,0000	0,0000	0,1000	0,2111
33	0,0000	0,0000	0,1000	0,0000	0,0714	0,0000	0,1714
34	0,1111	0,0000	0,0000	0,0000	0,0000	0,2500	0,3611
35	0,1000	0,2500	0,0769	0,0000	-0,3333	0,2857	0,3793
36	0,0000	0,1429	0,0625	0,0000	0,0000	0,1667	0,3720
37	0,3333	0,0000	0,2500	0,3333	-1,0000	0,0000	-0,0833
38	0,0000	0,1111	0,0000	0,0000	0,1000	0,0000	0,2111
39	0,0714	1,0000	0,1429	0,2000	0,0000	0,1667	1,5810
40	0,0000	0,0000	-0,4286	0,0000	0,0000	0,0000	-0,4286
41	0,0000	0,0000	0,0000	0,0000	-1,0000	0,2000	-0,8000
42	0,0000	-0,3333	0,0000	0,0000	-1,0000	0,0000	-1,3333
43	0,0000	-1,0000	0,3333	0,0000	0,0000	0,0000	-0,6667
44	0,0000	0,0000	0,1111	0,1000	0,0000	0,0000	0,2111
45	0,0000	0,5000	0,0000	0,0000	0,0000	-0,0476	0,4524
46	0,2500	0,0000	0,0980	0,0000	-0,0323	0,0000	0,3158
47	0,1250	1,0000	0,2500	0,0000	-1,0000	0,0000	0,3750
48	0,0000	0,0000	0,3333	0,5000	0,0000	0,0769	0,9103
49	0,0000	0,0000	0,1250	0,2000	-0,1000	0,2500	0,4750
50	0,0000	0,0000	0,6667	0,0000	0,1429	0,1667	0,9762
51	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
52	0,3333	0,0000	0,0000	0,3333	0,0000	0,0000	0,6667
53	0,0000	1,0000	0,0000	0,0000	-1,0000	-1,0000	-1,0000
54	0,0000	-1,0000	0,0000	0,5000	-1,0000	0,0000	-1,5000
55	0,0000	-0,2500	0,0000	0,0000	0,0000	0,6667	0,4167
56	0,2857	-1,0000	0,0000	-0,3333	0,2857	-0,5000	-1,2619
57	0,0909	0,0526	0,2500	0,1429	-0,1111	0,0000	0,4253
58	0,0000	0,5000	0,5000	0,1250	0,0000	0,0000	1,1250
59	0,0000	0,0000	0,0000	0,1429	0,0000	-0,1667	-0,0238
60	0,1250	0,0000	0,0000	0,2000	0,0000	0,0000	0,3250
61	0,4000	0,0000	0,2500	0,2727	-0,1250	0,0556	0,8533
62	0,0625	-0,0588	0,0000	0,0000	0,0000	0,0000	0,0037
63	0,1111	-0,0606	0,0625	0,0909	0,0000	-1,0000	-0,7961
64	0,0000	-1,0000	0,0000	0,0000	0,0000	0,1739	-0,8261
65	0,1429	-0,3333	0,0000	0,0000	0,0000	0,0000	-0,1905
66	0,0000	0,0000	-0,2500	0,0000	0,0000	0,3333	0,0833
67	0,3333	0,0000	0,0000	0,0000	0,0000	0,0000	0,3333
68	-0,1250	0,0000	0,0000	0,0000	1,0000	0,0000	0,8750
69	0,0000	-1,0000	0,0000	0,0000	-1,0000	0,0000	-2,0000

70	0,0000	1,0000	0,3333	1,0000	0,0000	0,0000	2,3333
71	0,2857	0,0000	0,0000	0,2000	0,0000	0,0000	0,4857
72	0,3333	-1,0000	-0,2000	1,0000	0,0000	0,3333	0,4667
73	0,1667	0,0000	0,1667	0,0000	0,0000	0,1429	0,4762
74	0,0000	-0,3333	0,1538	0,1429	0,0000	0,0000	-0,0366
75	0,1250	-1,0000	0,1333	-0,0500	0,0000	-0,3333	-1,1250
76	0,0000	0,0000	0,0000	0,0000	0,1667	0,0000	0,1667
77	0,1429	0,2500	0,0000	0,0000	0,0000	0,0000	0,3929
78	0,0000	0,2222	0,0000	0,0000	0,3333	0,6667	1,2222
79	0,5000	1,0000	-0,5000	0,0000	-1,0000	0,3333	0,3333
80	0,2500	-0,3333	0,0000	0,0000	0,0000	1,0000	0,9167
81	0,0000	1,0000	0,0000	0,0000	-1,0000	0,5000	0,5000
82	0,1667	0,0000	0,0000	0,0000	0,0000	0,0000	0,1667
83	0,5000	-1,0000	0,0000	0,0000	0,0000	-0,1429	-0,6429
84	0,1429	0,0000	0,2500	0,1333	-0,2500	0,1429	0,4190
85	0,0000	0,0000	0,0000	0,0000	-0,5000	0,1429	-0,3571
86	0,1000	0,3333	0,0000	0,0000	0,0000	0,2500	0,6833
87	0,0000	0,0000	0,0000	0,0000	-1,0000	0,0000	-1,0000
88	-0,0909	0,0000	-0,2500	0,3333	0,0000	-0,5000	-0,5076
89	1,0000	1,0000	0,3333	0,0000	0,0000	0,1905	2,5238
90	0,1429	0,0909	0,0833	0,0000	-0,0417	0,0000	0,2754
91	0,0000	0,0000	0,0000	0,5000	-0,3333	0,0000	0,1667
92	0,0000	-1,0000	0,3333	0,0000	0,0000	0,0000	-0,6667
93	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
94	0,5000	0,0000	0,0526	0,0000	0,1429	0,0000	0,6955
95	0,1429	0,0000	0,0000	0,2500	0,0000	0,0000	0,3929
96	0,0000	-0,3333	0,3333	0,0000	0,0000	0,0000	0,0000
97	0,0000	0,0000	0,0000	0,2500	0,0000	0,0000	0,2500
98	1,0000	0,0000	0,0000	0,3333	1,0000	1,0000	3,3333
99	0,3333	0,0000	0,5000	0,0000	0,0000	0,0000	0,8333
100	0,0000	0,0000	0,0000	1,0000	-1,0000	0,0000	0,0000
101	0,0588	-0,1250	0,1250	0,0000	0,0000	0,0000	0,0588
102	0,0000	0,0000	0,0000	0,1667	0,0000	0,0000	0,1667
103	0,0833	-0,2500	-0,0625	0,0667	0,2000	0,1429	0,1804
104	0,2500	0,1111	0,2000	0,1111	0,1000	0,0000	0,7722
105	0,2500	0,0000	0,0000	0,0000	0,0000	0,0625	0,3125
106	0,0000	0,1111	-0,0833	0,0909	0,0000	0,0588	0,1775
107	0,1429	0,0667	0,3333	0,0909	-0,0455	-0,1429	0,4455
108	0,0833	0,0000	0,0000	0,0909	0,0000	0,0000	0,1742

109	0,3333	-0,1667	0,3333	0,0000	-1,0000	0,0000	-0,5000
110	0,0000	0,0000	0,0000	0,2000	0,0000	0,2222	0,4222
111	0,1429	0,2000	0,5000	0,0000	-0,1000	0,0000	0,7429
112	0,1000	0,0000	0,0000	0,2857	0,0000	-0,5000	-0,1143
113	0,3333	1,0000	0,2000	0,3333	-0,1667	0,5000	2,2000
114	0,0000	0,0000	0,0000	0,0000	0,0000	-0,1000	-0,1000
115	0,2500	0,0000	0,0000	0,0000	-1,0000	0,0000	-0,7500
116	0,1250	0,0000	0,0000	0,0000	0,0000	0,0000	0,1250
117	0,6667	-1,0000	0,0000	0,0000	0,0000	0,0000	-0,3333
118	0,2143	0,0000	0,5000	0,5000	-0,2500	0,1667	1,1310
119	0,3333	0,2500	0,1250	0,0000	-0,2000	0,0000	0,5083
120	0,0000	0,3333	0,2000	0,0000	0,0000	0,3000	0,8333
121	-0,0909	0,0000	0,1667	0,0000	0,2000	-0,1250	0,1508
122	0,3000	0,0833	0,2000	0,0000	0,0000	0,0000	0,5833
123	0,0000	0,0000	0,3333	0,0000	0,0000	0,2000	0,5333
124	0,0000	0,3333	0,0000	0,5000	-1,0000	0,0000	-0,1667
125	0,0000	-0,1429	0,2500	0,0000	-1,0000	-0,5000	-1,3929
126	0,0909	0,0000	-0,1000	0,3333	-1,0000	0,3333	-0,3424
127	0,0000	-1,0000	0,0000	0,0000	0,0000	0,0000	-1,0000
128	0,1818	0,0000	0,0909	0,5000	-0,6667	0,0000	0,1061
129	0,0769	0,0286	0,0000	0,1429	0,0645	0,0000	0,3129
130	0,0000	0,0000	-0,1818	0,0000	0,0000	0,0000	-0,1818
131	0,2000	1,0000	0,0000	0,0000	-1,0000	0,0000	0,2000
Sum	14,220	-2,3738	10,466	16,908	-15,7136	8,2216	31,729

Table 5.11: Sentiment score of each respondent's answer to each question.

It is of note that negatively scored answers are the minority. This is due to the fact there is a positive bias in human language (Augustine, Mehl, & Larsen, 2011; Dodds et al., 2015). This makes the fact that the aggregate scores regarding questions two and five are negative an important indicator that the issues discussed in the questions is aggravating towards the population. In table 5.10 (even in questions two and five, which overall have a negative sentiment score), the fraction of positively scored answers is higher than negative ones.

A good way of looking for patterns here is to produce an histogram of the sentiment scores, these are produced for the answers to each question and presented in the following figures.

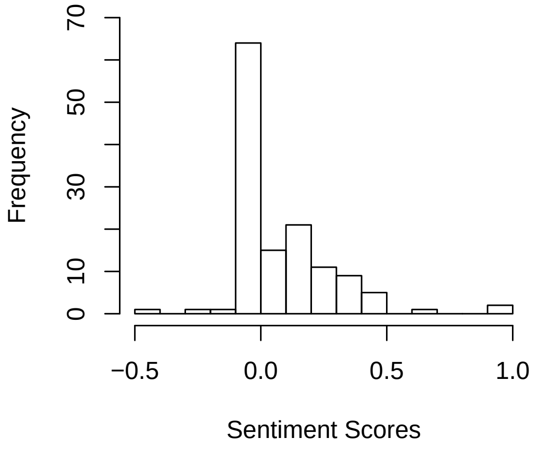


Figure 5.1: Histogram of the sentiment scores for the answers to question 1.

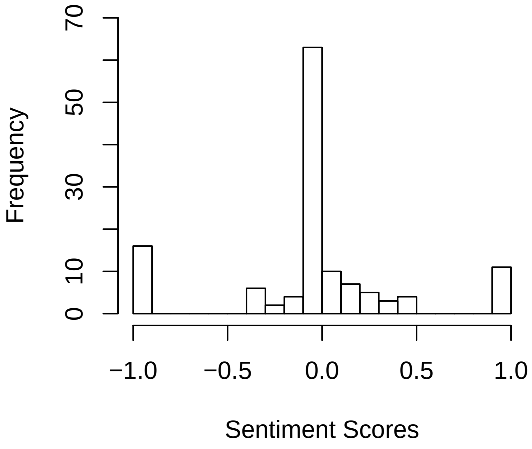


Figure 5.2: Histogram of the sentiment scores for the answers to question 2.

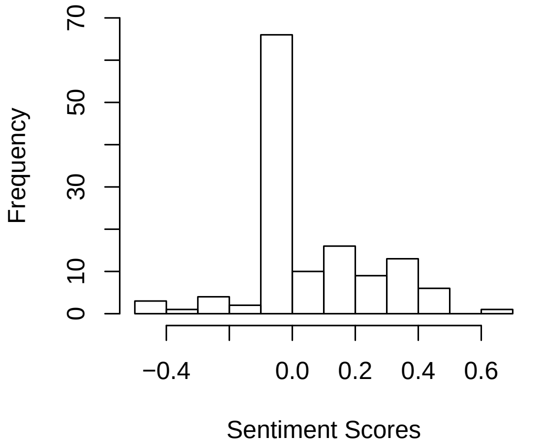


Figure 5.3: Histogram of the sentiment scores for the answers to question 3.

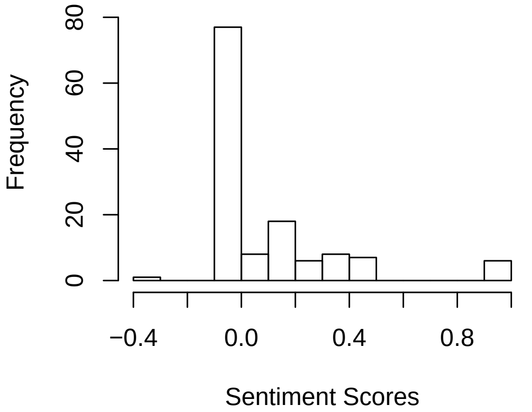


Figure 5.4: Histogram of the sentiment scores for the answers to question 4.

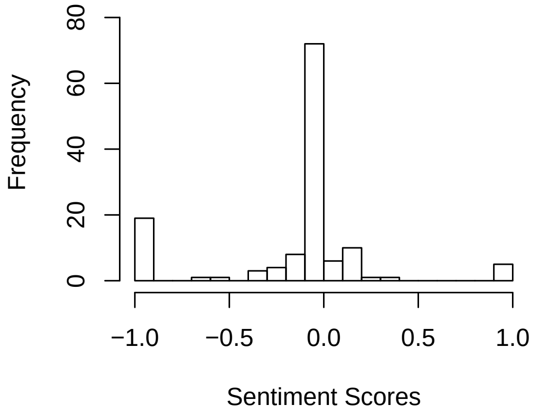


Figure 5.5: Histogram of the sentiment scores for the answers to question 5.

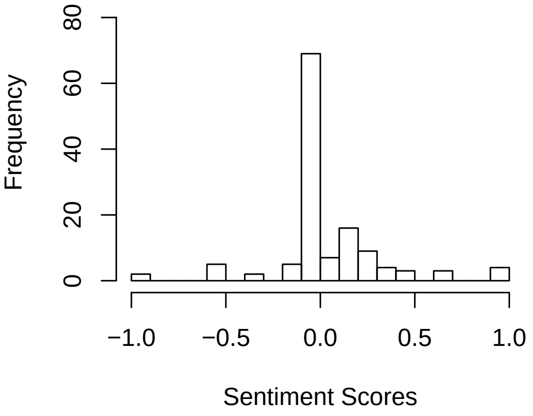


Figure 5.6: Histogram of the sentiment scores for the answers to question 6.

The collected data will be discussed in the Data Analysis section. However as a preliminary observation we can see that, in line with the cluster analysis, there is one bin, which contains more elements than all others. This hints that there are common or dominant characteristics present in IST's students.

Cluster Analysis

The clusters were obtained by applying a K-means clustering algorithm, to a matrix of document distances, computed using the cosine distance.

The number of clusters was set to four when analysing questions: 1, 2, 3 and 6. For questions 4 and 5 this was set to three since with four clusters the quality of the clustering would suffer. For example, the clustering of question four with four clusters had three empty clusters and one with all the documents. These numbers are chosen to be in line with the number of personas that we wish to generate.

The clusters are then named, manually, after inspecting the documents contained within it. Clusters of documents are related to personas, since students which can be described by the same persona have similar goals, beliefs and objectives, similar answers are expected when asked about those similar characteristics.

The clusters are presented here as tables which present the names and sizes of each cluster, for each question. Furthermore a table containing each individual entry of each cluster in each answer is provided in Appendix A as well as graphical representations of each cluster.

The attributed names and sizes each cluster are reported in the following tables.

Cluster	Name	Size
1	'scholars'	12
2	'wants'	15
3	NA	3
4	'future'	101

Table 5.12: Size of the clusters, question one answers.

Cluster	Name	Size
1	'happy, frustrated'	101
2	'mixed feelings'	11
3	'anxious, tired'	12
4	'frustated"	17

Table 5.13: Size of the clusters, question two answers.

Cluster	Name	Size
1	'achieve'	98
2	'improve'	21
3	'understand'	8
4	'autonomy'	4

Table 5.14: Size of the clusters, question three answers.

Cluster	Name	Size
1	'stability'	9
2	'happiness'	118
3	'focus, excellence'	4

Table 5.15: Size of the clusters, question four answers.

Cluster	Name	Size
1	'sad'	10
2	'anxious'	9
3	'worried, hopefull'	112

Table 5.16: Size of the clusters, question five answers.

Cluster	Name	Size
1	' - '	6
2	'back to normal'	104
3	' - '	6
4	' - '	13

Table 5.17: Size of the clusters, question six answers.

Topic Models

The function *'LDA'* from the package *'topicmodels'* is used. The number of topics is chosen empirically with the help of both the other analysis performed and the package *'ldatuning'* which generates quality measures for the available topics. The results of the tuning algorithm are shown in Appendix B. Topics are represented as a list of words that have a higher probability of belonging to a particular topic, proportional to the β quantity calculated by the topic model.

The topics allow us to discerne the overall issues being brought up in the answers, analysing individual answers with this information also informs us wether each respondent is more focused on the same or different issues when compared to the others. This information is important in describing a persona as it is connected with how the persona interprets the world.

The top 10 terms associated with each of the four topics is presented next in tables. Each group of four tables relates to one question. Question one topics are presented in the following tables.
It is to be noted that some topics have high probability terms that coincide with the cluster analysis. This makes those terms more relevant in the analysis.

term	beta
job	0,0472
skills	0,0158
education	0,0183
field	0,0107
degree	0,0176
future	0,0172
skills	0,0158
life	0,0142
study	0,0139
continue	0,0138

Table 5.18: Top ten terms regarding beta, question 1. Topic 1.

term	beta
learn	0,0746
degree	0,0283
job	0,0260
future	0,0244
education	0,0244
engineer	0,0177
science	0,0106
chances	0,0071
follow	0,0071
learn	0,0071

Table 5.19: Top ten terms regarding beta, question 1. Topic 2.

term	beta
knowledge	0,0228
moment	0,0114
step	0,0114
path	0,0114
degree	0,0113
education	0,0057
understanding	0,0057
social	0,0057
curious	0,0057
best	0,0057

Table 5.20: Top ten terms regarding beta, question 1. Topic 3.

term	beta
future	0,0182
university	0,0180
engineering	0,0170
learning	0,0170
enjoy	0,0170
feel	0,0127
wanted	0,0127
doors	0,0127
attending	0,0085
job	0,0085

Table 5.21: Top ten terms regarding beta, question 1. Topic 4.

The topics for the remaining answers to question are presented in the following tables.

term	beta
proud	0,1051
university	0,0166
nostalgic	0,0166
times	0,0147
happy	0,0110
mixed	0,0110
moments	0,0110
feelings	0,0110
hate	0,0110
accomplished	0,0055

Table 5.22: Top ten terms regarding beta, question 2. Topic 1.

term	beta
tired	0,0435
engineering	0,0272
depressed	0,0272
excited	0,0217
happy	0,0217
learn	0,0163
happy	0,0141
contempt	0,0108
degree	0,0108
dread	0,0108

Table 5.24: Top ten terms regarding beta, question 2. Topic 3.

term	beta
frustrated	0,0141
home	0,0141
panic	0,0141
smart	0,0141
stress	0,0141
anxious	0,0141
proud	0,0070
skills	0,0070
people	0,0070
accomplished	0,0070

Table 5.23: Top ten terms regarding beta, question 2. Topic 2.

term	beta
anxious	0,0651
happy	0,0548
stressed	0,0522
sad	0,0456
time	0,0261
university	0,0195
times	0,0152
feel	0,0130
education	0,0130
top	0,0130

Table 5.25: Top ten terms regarding beta, question 2. Topic 4.

term	beta
learn	0,0482
job	0,0289
career	0,0192
professional	0,0144
situations	0,0144
field	0,0096
achieve	0,0096
grow	0,0096
jobs	0,0096
easily	0,0096

Table 5.26: Top ten terms regarding beta, question 3. Topic 1.

term	beta
solve	0,0444
understand	0,0246
learn	0,0246
pursue	0,0197
lot	0,0148
future	0,0148
world	0,0148
prepared	0,0148
raising	0,0098
tecnico	0,0098

Table 5.28: Top ten terms regarding beta, question 3. Topic 3.

term	beta
overcome	0,0200
life	0,0100
better	0,0100
perspectives	0,0100
problem	0,0100
smart	0,0100
thrive	0,0100
critically	0,0100
challenge	0,0096
related	0,0050

Table 5.27: Top ten terms regarding beta, question 3. Topic 2.

term	beta
skills	0,0379
develop	0,0284
engineer	0,0142
faster	0,0142
reasoning	0,0095
soft	0,0094
teach	0,0094
solving	0,0094
fast	0,0094
pressure	0,0094

Table 5.29: Top ten terms regarding beta, question 3. Topic 4.

term	beta
develop	0,0194
feel	0,0180
projects	0,0129
creative	0,0129
enjoy	0,0129
flexible	0,0129
ideas	0,0129
learning	0,0129
specific	0,0129
professional	0,0064

Table 5.30: Top ten terms regarding beta, question 4. Topic 1.

term	beta
feel	0,0506
money	0,0248
learn	0,0217
money	0,0186
earn	0,0161
happy	0,0124
respected	0,0124
valued	0,0062
time	0,0062
balance	0,0062

Table 5.32: Top ten terms regarding beta, question 4. Topic 3.

term	beta
happy	0,0573
excel	0,0208
solve	0,0208
money	0,0208
grow	0,0156
environment	0,0108
feel	0,0107
motivated	0,0104
lot	0,0104
expectations	0,0104

Table 5.31: Top ten terms regarding beta, question 4. Topic 2.

term	beta
environment	0,0250
challenged	0,0193
lives	0,0193
contribute	0,0129
change	0,0129
constantly	0,0129
difference	0,0129
manage	0,0129
people	0,0129
real	0,0129

Table 5.33: Top ten terms regarding beta, question 4. Topic 4.

term	beta
anxious	0,0650
worried	0,0473
stressed	0,0295
hopeful	0,0177
anxiety	0,0118
life	0,0118
angry	0,0118
change	0,0118
due	0,0059
lives	0,0059

Table 5.34: Top ten terms regarding beta, question 5. Topic 1.

term	beta
future	0,0382
frustrated	0,0239
normal	0,0191
uncertain	0,0191
bored	0,0143
angry	0,0095
people	0,0095
life	0,0095
fearful	0,0095
resolve	0,0095

Table 5.36: Top ten terms regarding beta, question 5. Topic 3.

term	beta
scared	0,0229
time	0,0171
day	0,0114
difficult	0,0114
optimistic	0,0114
depressed	0,0057
anxious	0,0057
friends	0,0057
human	0,0057
life	0,0057

Table 5.35: Top ten terms regarding beta, question 5. Topic 2.

term	beta
sad	0,0986
depressed	0,0273
tired	0,0219
lives	0,0109
afraid	0,0109
nervous	0,0109
powerless	0,0109
social	0,0109
times	0,0060
event	0,0054

Table 5.37: Top ten terms regarding beta, question 5. Topic 4.

term	beta
explore	0,0249
world	0,0249
ist	0,0187
people	0,0177
everyday	0,0124
kill	0,0124
smoke	0,0124
thesis	0,0124
things	0,0124
simple	0,0124

Table 5.38: Top ten terms regarding beta, question 6. Topic 1.

term	beta
live	0,0323
change	0,0215
job	0,0212
study	0,0161
future	0,0107
house	0,0107
backyard	0,0107
mental	0,0107
skip	0,0107
quit	0,0107

Table 5.40: Top ten terms regarding beta, question 6. Topic 3.

term	beta
enjoy	0,0629
friends	0,0436
time	0,0339
life	0,0290
family	0,0193
skills	0,0145
university	0,0145
improve	0,0145
people	0,0104
difference	0,0096

Table 5.39: Top ten terms regarding beta, question 6. Topic 2.

term	beta
home	0,0273
leave	0,0218
travel	0,0218
house	0,0164
rethink	0,0109
related	0,0109
won	0,0109
stay	0,0109
quit	0,0054
health	0,0054

Table 5.41: Top ten terms regarding beta, question 6. Topic 4.

5.4 Data Analysis

Analysing ranked lists

From these lists some high level knowledge is already observable. For instance we can see that in Table 5.4, relating to question I, one of the highest ranking words is *'supposed'*, hinting, for instance, that some people attend college because it is expected of them. Still analysing question one, a couple of other words to keep in mind are *'ambitions', 'engineer', 'graduate'*. These point towards the goals of the respondents.

Question II is heavily loaded with sentiment by design, allowing the understanding of their mindset regarding their current environment. The list in Table 5.5 is very biased towards negative sentiment, for example *'stress', 'contempt'*. This is an interesting venue to pursue later when performing sentiment analysis. There are also participants which display positive sentiment here, for example *'smart', 'motivated'*, already allowing for a crucial differentiation in how the participants view the word.

Question III, which is loaded towards the future aims to discover the aspirations of the participants. In Table 5.6 *'expertise', 'grow', 'trust'* point towards a desire of acquiring reputation in a field, quite possibly *'research'*.

Also directed towards the future, question IV is more specific, in that it elicits knowledge specifically towards aspirations in the workspace. The terms are listed in Table 5.7. A clear separation is visible, *'dominate','thrive'* and *'confortable', 'happy'* are possibly considered forms of success by different respondents.

Question V relates the students with their current global environment. The list in Table 5.8 is also very negatively loaded, which is to be expected due to the current global pandemic.

In the last question information is collected on how each participant responds to the adverse conditions currently in play. The list is in Table 5.9. There is a clear desire for self improvement in the face of adversity.

Sentiment Analysis

Analysing the histograms we can see that the largest bin is for all questions is slightly below zero, this is because this bin also contains 0, which represents a Neutral score.

From table 5.10 we can see that Neutral is the most common class in almost all answers to all questions. From the histograms we can also observe that extremely Negative sentiment scores arise only in the answers to questions II and V, confirming the previous finding that these were the most negatively polarising questions. Furthermore, the other questions share a similar distribution between themselves. The positive bias found in these, can be confirmed in the last row of table 5.11.

Furthermore, we can take the average $\frac{31,729}{131} = 0,242$. Then comparing this with the value in the last column, for each respondent, we can check to see whether each respondent has a generally more Positive or Negative outlook when compared to his peers. This analysis is presented in the following table.

Higher	68
Lower	63
Sum	131

Table 5.42: Table representing whether the respondent had a higher or lower average sentiment score, compared to the average of all respondents.

This information is important due to the clear separation that arises between the beliefs of a respondent who scores higher than the average compared to one which scores lower.

Cluster Analysis

A first glance quickly gives us important information, indicating that the clusters are very unbalanced regarding membership numbers. There is one cluster which contains the majority of the documents in every case. Further analysis of the documents in each cluster will shed more light as to why this is the case.

Analysing the documents contained in each cluster should produce an aggregate of the responses that used similar terms, and per the distributional hypothesis mentioned in the Literature Review, convey similar meanings. As the questions were designed to collect information about proxies for the very basic characteristics of a persona these clusters are aggregating these different characteristics. A more detailed explanation is provided in the Data and Questionnaire section.

In Table 5.12 the third cluster is labeled NA as it contained three identical documents (same single word answer).

Looking at these tables we can summarily determine some of the goals in each group of clustered answers. Furthermore, the fact that there is an unbalanced distribution regarding cluster membership could point us towards the fact that some goals are considered important by a majority of the respondents while others are important to specific sub groups within them, allowing the separation of respondents according to these goals.

From table 5.12 there are 3 relevant clusters. Cluster 1 contains answers relating with a desire to learn and even pursue a career in research while cluster 4 has answers more geared towards securing a stable job in the future. Also in cluster 4 is where the 'supposed' previously mentioned in the ranked list analysis arises.

Moving on to the clusters of table 5.13, relating to the answer to question II, clusters 1-2 are related to answers which have mixed feelings regarding the institution. Clusters 3 and 4 are comprised of mostly negative answers about the institution, which reveals a trend, while most respondents have negative feelings towards the institution a majority of them also has positive ones, creating a clear avenue to separate respondents.

In table 5.14 there is a major focus on achieving their goals, represented by the cluster 1 in this table. However, in clusters 2 and 3 we see a desire to improve oneself and understand the world more deeply, without mentioning the achievement part of attending university, pointing to different goals and beliefs.

In the clusters from the answers to question IV an interesting separation emerges. The great majority of respondents stated they wanted to be happy. Then, two minority stances appear, some students claim to seek for stability while others strive for excellence at their field. Clearly the goals and beliefs of the respondents are different here, revealing some avenue of separation for the step of designing personas.

On the answers to question V, there are not any clearly contradicting clusters of opinions, all the respondents were somewhat apprehensive with the current situation. Also to note that the majority of them were also hopeful, not just worried about it.

Regarding the final clustering performed, in table 5.17, the results where not very conclusive. All the answers expressed a desire to go back to a normal routine, without a clear separation. The fact that most documents belong to one cluster makes sense here, however when checking the documents found in the other clusters we find that they express similar opinions and were clustered together due to other reasons, one of them being typos, for example.

Topic models analysis

On the topics derived from the answers to question 1, we can identify that topic 1 and 2 are clearly more focused towards finding a job. Whereas topics 3 and 4 are more related with learning to improve oneself or to satisfy curiosity. This is in line with what was found in the cluster analysis for this question, giving us a clear separation of respondents in this question. A good starting point towards building the personas.

Regarding the answers to question 2, here it is not really possible to separate between topics with ease. Topic 1 is clearly the most positive of the list, with some degree of mixed feelings (low ranking negative word is present). However, all topics are comprised of mixed words of positive and negative sentiment. This is also very much in line with the clustering performed previously, which revealed most student indeed have mixed feelings regarding the institution.

In the answers to question 3 we find some really interesting topics. Topic 1 is really focused on career aspects as one can see from the plethora of job related terms. Contrasting in Topics 3 and 4 the terms point us towards a desire to understand and learn more. These are clearly different goals, which will be exploited further in this work.

Looking now towards Topic 2 we see terms related with excelling and being better, but not necessarily in a particular area, which shows there is a group of students who strive towards being the best at what they do.

Analysing the topics of the answers to question 4 Topics 2 and 3 are similar, in which they talk about the goals for a future workplace, and both give a somewhat high degree of importance towards money and happiness, which are presumably the focus of some of the respondents.

Topics 1 and 4 on the other hand are more related towards goals of independence and contribution towards a bigger goal, which are very different ways of facing the prospect of a future workspace.

Question 5, which along with question 2, was one of the most negatively charged in sentiment, suffers from the same lack of separation. All the topics are essentially comprised of negatively charged sentiment words, which just shows that almost all respondents are facing this crisis in a similar way.

Topics for question 6 are not very conclusive, this is due to very similar answers from the part of everybody, mixed in with a few very distinct answers.

5.5 Persona Design

With the information collected and sorted, a deeper dive can begin. In the following section 3 personas will be detailed. They will be presented in the form of a table with the following entries: Goals, Motivators and Sentiment (with respect to IST). On top of this table, each persona will be finished with a short fictional bio to help with emphasising and understanding said persona. All names are fictional.

Upon analysis of the acquired data, 3 personas pop into mind, which are, in no particular order: 'The Scholar', 'The Dominator' and 'Stable Job'

Firstly, the 'Scholar' persona is presented.

Goals:	Motivators:	Sentiment:
Study	Understanding	Highly Positive
Learn	Curiosity	
Research	Knowledge	

Table 5.43: 'Scholar' persona.

They highly value knowledge and learning new skills, which, as will be seen is a common theme. However the 'Scholars' value this out of curiosity and a desire to understand the world, not necessarily as a means to an end but as a goal itself.

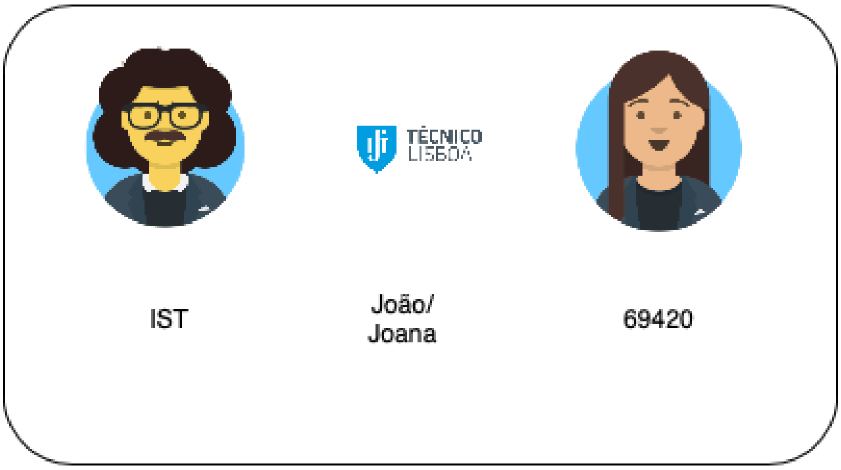


Figure 5.7: 'Scholar' persona student ID card.

The bio for this persona is:

```
He/She is a second year student. He/She is highly enthusiastic regarding the learning
environment at IST. He/She enjoys learning and is considering following a career in research.
He/She likes attending conferences regarding his areas of study.
```

The 'dominator' persona is related to people which want to be the best, both for the sake of being the best and for the rewards like status and money.

Goals:	Motivators:	Sentiment:
Excel	Competitive	Mostly Neutral
Overcome	Status	
Win	Money	

Table 5.44: ’Dominator’ persona.

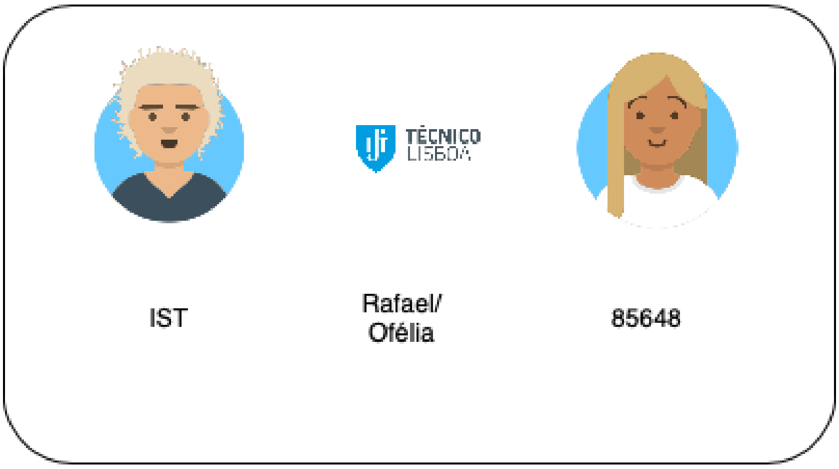


Figure 5.8: ’Dominator’ persona student ID card.

The bio for this persona is:

He/She is a fourth year student. He has had excellent grades so far.

He/She focuses on coursework as a means of obtaining a more favourable position in the future.

He/She aims for upper management positions.

Lastly we have ’Stable Job’, which was the most common type of persona found among the respondents. They seek stability and comfort in their daily lives.

Goals:	Motivators:	Sentiment:
Happiness	Money	Slightly negative
Stability	Societal Pressures	
Comfort	Duty	

Table 5.45: ’Stable Job’ persona.

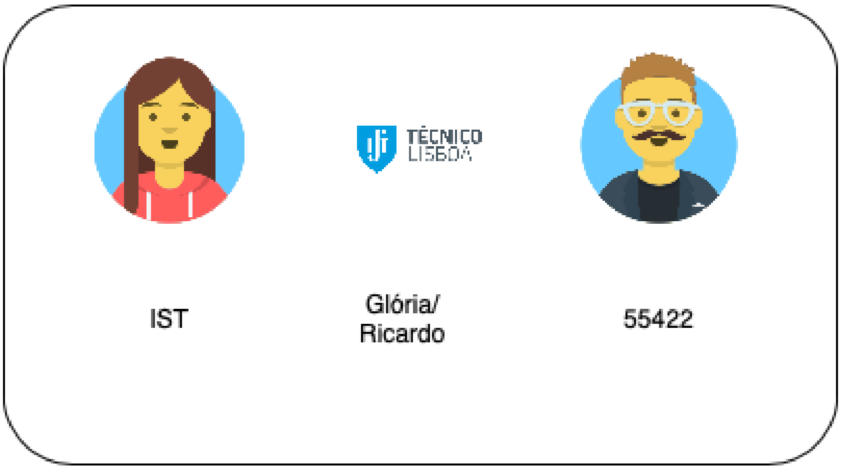


Figure 5.9: 'Stable Job' persona student ID card.

The bio for this persona is:

```
He/She is a first year student. He/She enrolled college as a means of obtaining a
stable job, by the influence of her parents. He/She dislikes the heavy coursework
 and time load of her course.
```

In the following chapter some considerations regarding these personas and their fit with the current student base of IST will be made. Also some ideas regarding how to verify some of these hypothesis will be provided.

Furthermore some ideas for future work, either related with the overall objective of the work of creating a tool that facilitates the analysis of open ended questions, or specifically related to the case study about IST's students that was developed in this work.

Chapter 6

Discussion and Conclusions

In this chapter the final conclusions regarding this work are presented as well as limitations, both for the work done and of the methods selected. A section will be dedicated towards important considerations towards possible future work. Finally, the overall work and key points will be addressed in the closing remarks section.

6.1 Discussion

To contextualise this discussion, the main objectives of this work must be kept in mind. One of these is the development of a methodology for enhancing persona development methods with the aid of computational methods, and, the other is the testing of said methodology. This was accomplished by applying the developed methodology to the case study presented in this work.

Some issues appeared when analysing the data. In some of the questions the answers were too similar for some of the techniques to be relevant. Moreover, the limited size of the dataset might have exacerbated this problem, since more respondents could increase the range of answers obtained. This affected mostly the cluster and topic model techniques, since these are the most statistically involved techniques used.

Reading the answers given, the Topic models and Clusters found make sense and point towards some of the personality traits proposed. Moreover, a great deal of information was extracted from the sentiment analysis and, surprisingly, from the ranked lists. The effectiveness of the simplest technique used may be due to the small data set used for the case study.

Furthermore, the techniques used allowed for the differentiation of deeper traits of the respondents. This is very important as this information is critical for persona development methods, which this work aims to improve.

Sentiment analysis was a valuable technique in grasping the overall opinion of the students regarding each question, as well as informing us of the position of a specific respondent regarding the issue approached in the question. The results obtained by each technique are according to expectations. For instance, the expected bias for the sentiment analysis section is observable in the histograms of the scores, as aforementioned.

The methodology is simple and delivers results quickly, especially when compared to manual analysis of the same dataset. The longest operation performed is the calibration of topic model parameters which requires fitting several models with different parameters. The size of the dataset comes into play once again, since due to being small, computations are very inexpensive and quick. In order to choose the number of clusters a similar operation is done, being, however, much less computationally expensive due to the nature of the calculations required for each technique to work.

This is a point to be careful with if an extremely large dataset is to be used. For small datasets with documents generated from a small number of topics this evaluation takes polynomial time. This means that the running time is upper bounded by a polynomial expression related with the size of the input. However, after a certain increase in the number of topics that generate the documents this computation becomes NP-hard (non-deterministic polynomial-time hardness), meaning it is hard to calculate even how long the operation will take (Sontag & Roy, 2010). In the case of analysing answers to open ended questions, the number of topics is somewhat bounded by the question, which helps with this potential problem of using this technique, for this application.

These overall positive results are encouraging and give direction towards a more streamlined and reproducible workflow for the development of personas. While there is space left for future work, which will be discussed further in this chapter, the objective of creating a tool that can enhance current methods for the development of personas was achieved.

Comparing this method with the ones exposed in the literature review, the main difference is that in the present work multiple text analysis techniques are applied in order to gather information, whereas in past work only one technique was used. For instance, in (Thoma & Williams, 2009) only clustering is used as a text analysis tool. The fact that combining multiple text analysis techniques allowed for a deeper analysis of the information, while not increasing the analysis time significantly is a major contribution from this work towards the literature in this area.

Regarding the personas this is one of the first iterations in developing a reproducible and modifiable methodology for this task. Traditionally analysing free text requires multiple human readers, turning it into a costly process (Roberts et al., 2014).

The main advantages gained over tradicional methods are the speed and scope of the analysis, as well as the lower costs. Another big advantage is the reproducibility of the process, since the same code will yield the same results when applied to the same dataset. The exception being the topic models, they change every time they are evaluated due to the sampling methods used. They should, however, produce results similar enough between different runs that this effect can be safely ignored.

As a preliminary exploration of existing methods applied to a different problem domain, some difficulties were expected and, overall, the combination and application of these techniques was quite effective. Taking this into consideration, we can say that the objective of applying the tool in this case study was completed with success.

6.2 Limitations

In this section limitations regarding this work will be discussed. The discussion starts with discussing limitations regarding the application of the developed methodology.

The first issue that comes into play has to do with the sample, which is not well-balanced and therefore not representative of the overall population under study. In essence, the personas developed can not be used as an evaluation or decision tool regarding the students, since the sample is not representative of the population, due to its small size when compared to the universe of students at IST.
It should be noted, however, that these biases do not make impossible the task at hand of testing the persona design method proposed using real data. The trade off from using a data set with these limitations is that there are no statistical guarantees for any results obtained.
Taking into account the complexity of each method, it is not easy to estimate a minimum number that would guarantee statistical soundness from the results, as said limit is different for each technique applied. On top of this, these limits vary with the content of each document in the dataset, meaning that it is only possible to verify if the results are significant after performing the analysis.

As this was a starting point for this type of methodology, only unsupervised techniques could be applied. No annotated datasets for this specific task were available at the moment this document was prepared. This limitation is important, since supervised classification techniques have high potential for the grouping of respondents according to specific characteristics, which could be very useful in generating personas.

To continue, limitations with the design of the methodology are presented, as well as suggestions for how to deal with them.

In this work, the final step of analysing the resulting data from applying each technique was performed manually, which is a limitation towards achieving a fully automated workflow for the generation of personas. However, this is also a limitation regarding the whole scope of the work done, the needed interpretation to go from groups of students to personas is a step which is very hard to code. This could be achieved with the aforementioned supervised techniques. However, these would require a significant amount of work previously performed by humans, for each task, defeating the purpose for this application. Furthermore, these annotated datasets are very specific, meaning that this type of knowledge does not transfer well from one problem to another. For instance, classification of an annotated dataset of movie reviews translates poorly into analysing financial documents, since words can take on a very different meaning, depending on a sub-domain's context (Loughran & McDonald, 2011a).

The present work, while having some limitations, is a solid base for further extension and development. Some suggestions will be provided in the next section, regarding both the techniques applied as a part of the developed methodology and the case study.

6.3 Future work

The NLP field is constantly growing and creating better tools that can be fit in this modular analysis workflow. This allows for an amazing degree of extensibility and modification of this workflow for other types of textual data, such as Tweets or costumer reviews.

To begin this section a discussion regarding the methodological steps and further work to be done on these is presented.
The three major techniques applied to this work were: SA, Clustering and Topic Modelling.

Regarding the SA step there are several directions that can be taken for future improvements. For instance, the chosen dictionary could be upgraded with words specific to this dataset, manually. Still in line with these improvements, automatic methods for dictionary generation can be explored. There are methods for this that do not require annotated datasets such as described in (Ahmed, Chen, & Li, 2020; Feng, Gong, Li, & Lau, 2018). This is expected to improve SA performance, especially when considering a larger dataset.
Furthermore, techniques in which a neural network intrinsically learns a useful representation automatically without human efforts could be applied as in (Guan et al., 2016). These techniques are known as weakly supervised and have proven to be effective in various sentiment analysis tasks such as product, movie and other reviews (Kayal, Singh, & Goyal, 2019; Zhao et al., 2018). Some of these techniques are very promising and give results in line with the state of the art supervised learning techniques previously used (Zeng, Zhou, Liu, & Song, 2019).

The clustering step was performed at the document level, with each answer to each question being considered a document. This could be done differently, as, for instance a document could be taken to be the set of answers from each respondent. The results will differ with the change in the scope of the clustering, a more detailed study on how this change affects the clustering results is appropriate at this point. In this work the clusters were based on simple document distances derived from the *tf-idf* weighting scheme. As a starting point this is acceptable. However, there are more advanced textual clustering techniques that make use of better features taking into account polysemy and semantic relations between words such as (Li, Cai, & Wang, 2020), which makes use of BERT, a ML based technique to generate feature vectors. Another promising technique for this could be the one presented in (Yi, Zhang, Zhao, & Wan, 2017), which uses a vocabulary network generated from a deep-learning algorithm applied to word co-occurrence matrices. Future work on this problem should focus on these improved techniques, as they report better clustering performance.

To conclude, the future work suggestions related to the applied techniques Topic Models are discussed. The simplest Topic Model available was applied, LDA, which could be improved by using other models. For instance, correlated topic models, an improvement of LDA by the original author does not impose topic independence. The strong independence assumption imposed by the Dirichlet in LDA is not realistic when analysing document collections, where one may find strong correlations between topics (D. M. Blei & Lafferty, 2007). This is the case in this work since answers to the same question are expected to be somewhat related.

A good idea would be to test different generative models for the topics, as there are methods that do not make use of the statistical priors that force the assumptions onto the model. An example of this is found in (Gerlach, Peixoto, & Altmann, 2018). In this work the topics are derived from network graph representations of the text. Better results are claimed by the authors making it worthwhile to pursue as an avenue for future work.

To summarise, the available techniques are constantly being improved and implemented in different ways. This allows for a fast evolution of the present methodology by upgrading the used techniques and checking if the results improve. It may also be important to cross test the different methodologies available, to see which ones are a better fit with each other is an important task for the future.

Regarding the case study, verification work could be done. Another questionnaire, for instance, could be used to try to understand if the students identify with any of the proposed personas. Another method could be used to verify the obtained results would be the generation of personas using the traditional methods in order to have a baseline of what to expect. Another worthy pursuit would be the application of the methodology to another dataset of answers to an open ended questionnaire, to better understand the generality of the tool.

While these would be valuable to the present work, important and more general tasks can be pursued, aiming to improve the entire eco-system of available resources to deal with this task.

A database of answers to open ended questionnaires would be a valuable tool for work in this area. It would be even more valuable with annotations, as it would allow for supervised techniques to be used. Furthermore, the existence of a dataset of similar problems allows for transfer learning of the generated word representations (Houlsby et al., 2019; Stickland & Murray, 2019).
This is important as training these models is quite time consuming, even with appropriate hardware (You et al., 2020). Furthermore, some standardised processing pipelines to work with the database would be valuable as exploration tools and as a starting point for future work.

Overall there is a lot to explore in this field, which is currently undergoing rapid growth. This coupled with the fact that a lot of important information is expressed in free text format (for example, costumer reviews), creates a frontier where work can be developed as a way of adapting and improving the current tools available. This should be the overall focus of future work in this area.

6.4 Closing remarks

The tool, being a collection of R language scripts is easy to apply and extend to many types of input data sets and to generate different analysis,

This work explores an area with a lot of development on individual techniques, but not in the application of them as a whole, in order to extract higher level information. Several insights were gained regarding the respondents of the questionnaire, which reveals a successful application of the selected techniques. It also shows that the proposed analysis pipeline can be a valid methodology for analysing the answers of open ended questionnaires, The generated personas are sensible regarding to what was expected at the start of the work.

As this was a preliminar work in the field, there are a lot of suggestions for future work, in many different areas. However, overall, the application of the natural language techniques necessary was successful and allowed for the collection of interesting results.

As a whole, the work provided promising results and showed that it is possible to devise an analysis pipeline for analysing free text, in the form of answers to an open ended questionnaire, using existing computational textual analysis techniques.

References

Ahmed, M., Chen, Q., & Li, Z. (2020, March). Constructing domain-dependent sentiment dictionary for sentiment analysis. *Neural Computing and Applications*, *32*(18), 14719–14732.

Airoldi, E. M. (2014). *Handbook of mixed membership models and their applications*. Chapman and Hall/CRC.

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017, August). Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET).* IEEE.

Almeida, F., & Xexéo, G. (2019). Word embeddings: A survey. *CoRR*, *abs/1901.09069*.

Alsmadi, M., Omar, K., & Azman, N. (2009, 01). Back propagation algorithm : The best algorithm among the multi-layer perceptron algorithm. *International Journal of Computer Science and Network Security*, *9*, 378-383.

Anaıs, C., Crina, C., Damien, J., Omar, H., & Lionel, B. (2013). A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation. *Research report RR-LIRIS-2014-002*.

Arun, R., Suresh, V., Madhavan, C. E. V., & Murthy, M. N. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Advances in knowledge discovery and data mining* (pp. 391–402). Springer Berlin Heidelberg.

Augustine, A. A., Mehl, M. R., & Larsen, R. J. (2011, February). A positivity bias in written and spoken english and its moderation by personality and gender. *Social Psychological and Personality Science*, *2*(5), 508–515.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). *Don't Count, Predict! A systematic comparison of context-counting vs. context predicting semantic vectors.*

Blei, D. (2006). *Twenty-first national conference on artificial intelligence aaai-06/eighteenth innovation applications*. Amer Assn for Artificial.

Blei, D. M., & Lafferty, J. D. (2007, December). Correction: A correlated topic model of science. *The Annals of Applied Statistics*, *1*(2), 634–634.

Brent, M. R. (1997). Toward a unified model of lexical acquisition and lexical access. *Journal of Psycholinguistic Research*, *26*(3), 363–375.

Brogueira, P., Gonçalves, A. P., Prazeres, D. M., Marrucho, I., Bioucas, J., ao Pimentel Nunes, J., ... Nunes, N. J. (2019, February). *Relatório final.*

Brouse, S. H. (2002, March). Interpreting qualitative data: Methods for analysing talk, text and interaction, 2nd edition by david silverman. sage, london, 2001, 325 pages, f17.99, ISBN 0 761 96865 2.

Journal of Advanced Nursing, *37*(6), 607–607.

Burges, C. (1998, June). A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, *2*(2), 121–167.

Burgess, B. (1994). *Analyzing qualitative data*. Routledge.

Burnard, P., Gill, P., Stewart, K., Treasure, E., & Chadwick, B. (2008, April). Analysing and presenting qualitative data. *British Dental Journal*, *204*(8), 429–432.

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on machine learning* (pp. 161–168). New York, NY, USA: ACM.

Chang, J., Boyd-graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Proceedings of the 22nd international conference on neural information processing systems.*

Chapman, C. N., & Milham, R. P. (2006, October). The personas' new clothes: Methodological and practical arguments against a popular method. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(5), 634–636.

Cooper, A. (1999). *The inmates are running the asylum*. Indianapolis, IN, USA: Macmillan Publishing Co., Inc.

Cortes, C., & Vapnik, V. (1995, September). Support-vector networks. *Mach. Learn.*, *20*(3), 273–297.

Dalal, M. K., & Zaveri, M. A. (2014). Opinion mining from online user reviews using fuzzy linguistic hedges. *Applied Computational Intelligence and Soft Computing*, *2014*, 1–9.

Devika, M., Sunitha, C., & Ganesh, A. (2016). Sentiment analysis: A comparative study on different approaches. *Procedia Computer Science*, *87*, 44–49.

Dodds, P. S., Clark, E. M., Desu, S., Frank, M. R., Reagan, A. J., Williams, J. R., ... Danforth, C. M. (2015, February). Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, *112*(8), 2389–2394.

Doyle, G., & Elkan, C. (2009). Accounting for burstiness in topic models. In *Proceedings of the 26th annual international conference on machine learning - 09.* ACM Press.

E., R. M., M., S. B., & Dustin, T. (2016). Navigating the local modes of big data: The case of topic models. *Computational Social Science: Discovery and Prediction*, 51–97.

Eisenhardt, K. M. (1989, October). Building theories from case study research. *The Academy of Management Review*, *14*(4), 532.

Engwall, L. (1983, July). Research note: Linguistic analysis of an open-ended questionnaire in an organizational study. *Organization Studies*, *4*(3), 261–270.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in r. *Journal of Statistical Software, Articles*, *25*(5), 1–54.

Feldman, R. (2006). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge University Press.

Feng, J., Gong, C., Li, X., & Lau, R. Y. K. (2018, August). Automatic approach of sentiment lexicon generation for mobile shopping reviews. *Wireless Communications and Mobile Computing*, *2018*, 1–13.

Fergnani, A. (2019, August). The future persona: a futures method to let your scenarios come to life. *foresight*, *21*(4), 445–466.

Fielding, J., Fielding, N., & Hughes, G. (2012, May). Opening up open-ended survey data using qualitative software. *Quality & Quantity*, *47*(6), 3261–3276.

Geer, J. G. (1991). Do open-ended questions measure salient issues? *Public Opinion Quarterly*, *55*(3), 360.

Gerlach, M., Peixoto, T. P., & Altmann, E. G. (2018, July). A network approach to topic models. *Science Advances*, *4*(7).

Gilbert, J. R., Moler, C., & Schreiber, R. (1992, January). Sparse matrices in MATLAB: Design and implementation. *SIAM Journal on Matrix Analysis and Applications*, *13*(1), 333–356.

Giorgetti, D., Prodanof, I., & Sebastiani, F. (2003). Automatic coding of open-ended questions using text categorization techniques. In *Proceedings of the 4th international conference of the association for survey computing (ascic 2003)* (pp. 173–184).

Goodwin, K. (2009). *Designing for the digital age: How to create human-centred products and services.* Wiley.

Graneheim, U., & Lundman, B. (2004, February). Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. *Nurse Education Today*, *24*(2), 105–112.

Guan, Z., Chen, L., Zhao, W., Zheng, Y., Tan, S., & Cai, D. (2016). Weakly-supervised deep learning for customer review sentiment classification. In *Proceedings of the twenty-fifth international joint conference on artificial intelligence* (p. 3719–3725). AAAI Press.

Gupta, V., & Lehal, G. S. (2009, August). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1).

Haykin, S. (1998). *Neural networks: A comprehensive foundation* (2nd ed.). Upper Saddle River, NJ, USA: Prentice Hall PTR.

Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014, January). Word cloud explorer: Text analytics based on word clouds. In *2014 47th hawaii international conference on system sciences.* IEEE.

Heinrich, G. (2009). A generic approach to topic models. In *Proceedings of the 2009th european conference on machine learning and knowledge discovery in databases - volume part i* (pp. 517–532). Berlin, Heidelberg: Springer-Verlag.

Hill, C. G., Haag, M., Oleson, A., Mendez, C., Marsden, N., Sarma, A., & Burnett, M. (2017). Gender-inclusiveness personas vs. stereotyping. In *Proceedings of the 2017 conference on human factors in computing systems.* ACM Press.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., … Gelly, S. (2019). *Parameter-efficient transfer learning for nlp.*

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An introduction to statistical learning: With applications in r.* Springer Publishing Company, Incorporated.

Jordan, M. (1998, 09). *Why the logistic function? a tutorial discussion on probabilities and neural* *networks.*

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (1st ed.). Upper Saddle River, NJ, USA: Prentice Hall PTR.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002, July). An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 881-892.

Kayal, P., Singh, M., & Goyal, P. (2019). Weakly-supervised deep learning for domain invariant sentiment classification. *CoRR*, *abs/1910.13425*.

Kim, Y. (2014). *Convolutional neural networks for sentence classification* (Vol. abs/1408.5882).

King, B. E., & Reinold, K. (2008). Natural language processing. In *Finding the concept, not just the word* (pp. 67–78). Elsevier.

Kumar, A. (2016). *Mastering text mining with r.* Packt Publishing.

Kuo, B. Y.-L., Hentrich, T., Good, B. M. ., & Wilkinson, M. D. (2007). Tag clouds for summarizing web search results. In *Proceedings of the 16th international conference on world wide web - 07.* ACM Press.

L, T., Laham, D., Rehder, B., & Schreiner, M. (1999, 01). How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans..

Landauer, T. K., Foltz, P. W., & Laham, D. (1998, January). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2-3), 259–284.

Lavrakas, P. (2008). *Encyclopedia of survey research methods.* Sage Publications, Inc.

Leleu, T. (2008, 10). Evaluating responses to open-ended questionnaire fields using latent semantic analysis..

Leleu, T. D., Jacobson, I. G., Leardmann, C. A., Smith, B., Foltz, P., Amoroso, P. J., ... Smith, T. C. (2011). Application of latent semantic analysis for open-ended responses in a large, epidemiologic study. In *Bmc medical research methodology.*

Li, Y., Cai, J., & Wang, J. (2020). A text document clustering method based on weighted bert model. In *2020 ieee 4th information technology, networking, electronic and automation control conference (itnec)* (Vol. 1, p. 1426-1430).

Lilley, M., Pyper, A., & Attwood, S. (2012, June). Understanding the student experience through the use of personas. *Innovation in Teaching and Learning in Information and Computer Sciences*, *11*(1), 4–13.

Ljungberg, B. F. (2017). Dimensionality reduction for bag-of-words models : Pca vs lsa..

Looker, E., Denton, M. A., & Davis, C. K. (1989, December). Bridging the gap: Incorporating qualitative data into quantitative analyses. *Social Science Research*, *18*(4), 313–330.

Loughran, T., & McDonald, B. (2011a). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, *66*(1), 35-65.

Loughran, T., & McDonald, B. (2011b, January). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, *66*(1), 35–65.

Maalej, W., Kurtanović, Z., Nabil, H., & Stanik, C. (2016, May). On the automatic classification of app reviews. *Requirements Engineering*, *21*(3), 311–331.

Madhoushi, Z., Hamdan, A. R., & Zainudin, S. (2015, July). Sentiment analysis techniques in recent works. In *2015 science and information conference (SAI).* IEEE.

Maimon, O., & Rokach, L. (Eds.). (2005). *Data mining and knowledge discovery handbook*. Springer-Verlag.

Manning, C., Raghavan, P., & Schutze, H. (2008). *Introduction to information retrieval*. New York, NY, USA: Cambridge University Press.

Maron, M. E. (1961, July). Automatic indexing: An experimental inquiry. *J. ACM*, *8*(3), 404–417.

Marsland, S. (2014). *Machine learning: An algorithmic perspective, second edition* (2nd ed.). Chapman & Hall/CRC.

Martin, D. I., Martin, J. C., & Berry, M. W. (2016). The application of LSA to the evaluation of questionnaire responses. In *Unsupervised learning algorithms* (pp. 449–484). Springer International Publishing.

Mcdonald, S., & Ramscar, M. (2001). Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *In proceedings of the 23rd annual conference of the cognitive science society* (pp. 611–6).

McGinn, J. J., & Kotamraju, N. (2008). Data-driven persona development. In *Proceeding of the twenty-sixth annual conference on human factors in computing systems - '08.* ACM Press.

Miaskiewicz, T., Sumner, T., & Kozar, K. A. (2008). A latent semantic analysis methodology for the identification and creation of personas. In *Proceeding of the twenty-sixth annual conference on human factors in computing systems - '08.* ACM Press.

Mich, L., Franch, M., & Inverardi, P. N. (2004, Feb 01). Market research for requirements analysis using linguistic tools. *Requirements Engineering*, *9*(1), 40–56.

Mikolov, T., tau Yih, W., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Hlt-naacl* (pp. 746–751).

Miles, M. B. (1979, December). Qualitative data as an attractive nuisance: The problem of analysis. *Administrative Science Quarterly*, *24*(4), 590.

Mitchell, T. M. (1997). *Machine learning* (1st ed.). New York, NY, USA: McGraw-Hill, Inc.

Mossholder, K. W., Settoon, R. P., Harris, S. G., & Armenakis, A. A. (1995, April). Measuring emotion in open-ended survey responses: An application of textual data analysis. *Journal of Management*, *21*(2), 335–355.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. The MIT Press.

Nguyen, E. (2014). Text mining and network analysis of digital libraries in r. In *Data mining applications with r* (pp. 95–115). Elsevier.

Nielsen, F. Å. (2011). A new ANEW: evaluation of a word list for sentiment analysis in microblogs. *CoRR*, *abs/1103.2903*.

Nitin, I., & Fred, D. (2010). *Handbook of Natural Language Processing*.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In

Empirical methods in natural language processing (emnlp) (pp. 1532–1543).

Peter, T. (2008). The latent relation mapping engine: Algorithm and experiments. *Journal of Artificial Intelligence Research*, *33*, 615–655.

Pietsch, A.-S., & Lessmann, S. (2018, July). Topic modeling for analyzing open-ended survey responses. *Journal of Business Analytics*, *1*(2), 93–116.

Popping, R. (2015, September). Analyzing open-ended questions by means of text analysis procedures. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *128*(1), 23–39.

Powers, D. M. W. (1998). Applications and explanations of zipf's law. In *Proceedings of the joint conferences on new methods in language processing and computational natural language learning* (pp. 151–160). Stroudsburg, PA, USA: Association for Computational Linguistics.

Pruitt, J., & Adlin, T. (2005). *The persona lifecycle: Keeping people in mind throughout product design*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Pruitt, J., & Grudin, J. (2003). Personas: Practice and theory. In *Proceedings of the 2003 conference on designing for user experiences - '03.* ACM Press.

Rish, I. (2001, 01). An empirical study of the naïve bayes classifier. *IJCAI 2001 Work Empir Methods Artif Intell*, *3*.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... Rand, D. G. (2014, March). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082.

Rocha, M., Cortez, P., & Neves, J. (2007, October). Evolution of neural networks for classification and regression. *Neurocomput.*, *70*(16-18), 2809–2816.

Rokade, A., Patil, B., Rajani, S., Revandkar, S., & Shedge, R. (2018, April). Automated grading system using natural language processing. In *2018 second international conference on inventive communication and computational technologies (ICICCT).* IEEE.

Ronan, C., Jason, W., Léon, B., Michael, K., Koray, K., & Pavel, K. (2011, November). Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, *12*, 2493–2537.

Russell, S., & Norvig, P. (2009). *Artificial intelligence: A modern approach* (3rd ed.). Upper Saddle River, NJ, USA: Prentice Hall Press.

Sahu, L., & Mohan, B. R. (2014, December). An improved k-means algorithm using modified cosine distance measure for document clustering using mahout with hadoop. In *2014 9th international conference on industrial and information systems (ICIIS).* IEEE.

Salminen, J., Jansen, B. J., An, J., Kwak, H., & gyo Jung, S. (2018, November). Are personas done? evaluating their usefulness in the age of digital analytics. *Persona Studies*, *4*(2), 47.

Schmidhuber, J. (2014). Deep learning in neural networks: An overview. *CoRR*, *abs/1404.7828*.

Schmidt, M. (2010, July). Quantification of transcripts from depth interviews, open ended responses and focus groups: Challenges, accomplishments, new applications and perspectives for market research. *International Journal of Market Research*, *52*(4), 483–509.

Schofield, A., Magnusson, M., & Mimno, D. (2017). Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of the 15th conference of the european chapter of the association* *for computational linguistics: Volume 2, short papers.* Association for Computational Linguistics.

Schofield, A., Magnusson, M., Thompson, L., & Mimno, D. (2017). Pre-processing for latent dirichlet allocation..

Silge, J. (2017). *Text mining with r: A tidy approach.* O'Reilly Media.

Simon, A., Deo, M. S., Selvam, V., & Babu, R. (2016, 01). An overview of machine learning and its applications. *International Journal of Electrical Sciences & Engineering*, *Volume*, 22-24.

Singh, A., Thakur, N., & Sharma, A. (2016, March). A review of supervised machine learning algorithms. In *2016 3rd international conference on computing for sustainable global development (indiacom)* (p. 1310-1315).

Sontag, D., & Roy, D. (2010, 06). Complexity of inference in topic models.

Stasko, J., Gorg, C., Liu, Z., & Singhal, K. (2007, October). Jigsaw: Supporting investigative analysis through interactive visualization. In *2007 IEEE symposium on visual analytics science and technology.* IEEE.

Stickland, A. C., & Murray, I. (2019). *Bert and pals: Projected attention layers for efficient adaptation in multi-task learning.*

Tan, C.-M., Wang, Y.-F., & Lee, C.-D. (2002, July). The use of bigrams to enhance text categorization. *Information Processing & Management*, *38*(4), 529–546.

Thoma, V., & Williams, B. (2009, 08). Developing and validating personas in e-commerce: A heuristic approach. In (p. 524-527).

Tomas, M., Ilya, S., Kai, C., Greg, C., & Jeffrey, D. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th international conference on neural information processing systems - volume 2* (pp. 3111–3119). USA: Curran Associates Inc.

Turney, P. D., & Pantel, P. (2010, February). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Welbers, K., Atteveldt, W. V., & Benoit, K. (2017, October). Text analysis in r. *Communication Methods and Measures*, *11*(4), 245–265.

Wu, Y., Wei, F., Liu, S., Au, N., Cui, W., Zhou, H., & Qu, H. (2010, November). OpinionSeer: Interactive visualization of hotel customer feedback. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 1109–1118.

Yi, J., Zhang, Y., Zhao, X., & Wan, J. (2017). A novel text clustering approach using deep-learning vocabulary network. *Mathematical Problems in Engineering*, *2017*, 1–13.

Yong, A. G., & Pearce, S. (2013, October). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, *9*(2), 79–94.

You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., ... Hsieh, C.-J. (2020). *Large batch optimization for deep learning: Training bert in 76 minutes.*

Zellig, H. (1954). Distributional Structure. *WORD*, *10*(2-3), 146–162.

Zeng, Z., Shi, H., Wu, Y., & Hong, Z. (2015). Survey of natural language processing techniques in bioinformatics. *Computational and Mathematical Methods in Medicine*, *2015*, 1–10.

Zeng, Z., Zhou, W., Liu, X., & Song, Y. (2019, June). A variational approach to weakly supervised document-level multi-aspect sentiment classification. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 386–396). Minneapolis, Minnesota: Association for Computational Linguistics.

Zhang, L., & Liu, B. (2016). *Sentiment analysis and opinion mining*. Springer US.

Zhang, X., & LeCun, Y. (2015). Text understanding from scratch. *CoRR*, *abs/1502.01710*.

Zhang, Y., & Wallace, B. (2015). *A sensitivity analysis of (and practitioners guide to) convolutional neural networks for sentence classification.*

Zhao, W., Guan, Z., Chen, L., He, X., Cai, D., Wang, B., & Wang, Q. (2018, January). Weakly-supervised deep embedding for product review sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, *30*(1), 185–197.

Appendix A

Cluster assignments

In this appendix the individual cluster assignments are shown, as tables.

Table A.1: Cluster assignment for the answers to question one.

Document	Cluster
1	4
2	1
3	4
4	4
5	4
6	4
7	4
8	4
9	4
10	4
11	3
12	3
13	4
14	1
15	4
16	4
17	4
18	4
19	4
20	4
21	2
22	2
23	4

24	4
25	4
26	4
27	4
28	4
29	4
30	4
31	1
32	4
33	4
34	2
35	4
36	4
37	1
38	4
39	4
40	4
41	2
42	4
43	4
44	4
45	4
46	2
47	4
48	4
49	4
50	4
51	4
52	4
53	4
54	2
55	4
56	4
57	4
58	4
59	4
60	4
61	4
62	4

63	4
64	4
65	4
66	4
67	2
68	4
69	2
70	4
71	4
72	1
73	2
74	4
75	4
76	3
77	4
78	4
79	4
80	4
81	4
82	1
83	1
84	4
85	4
86	1
87	4
88	4
89	1
90	4
91	4
92	4
93	4
94	4
95	4
96	4
97	4
98	1
99	4
100	4
101	4

102	1
103	4
104	4
105	4
106	4
107	4
108	4
109	4
110	4
111	2
112	4
113	2
114	4
115	2
116	4
117	4
118	2
119	1
120	4
121	4
122	2
123	2
124	4
125	4
126	4
127	4
128	4
129	4
130	4
131	4

Table A.2: Cluster assignment for the answers to question two.

Document	Cluster
1	4
2	1
3	4
4	4

5	4
6	4
7	4
8	4
9	4
10	4
11	3
12	3
13	4
14	1
15	4
16	4
17	4
18	4
19	4
20	4
21	2
22	2
23	4
24	4
25	4
26	4
27	4
28	4
29	4
30	4
31	1
32	4
33	4
34	2
35	4
36	4
37	1
38	4
39	4
40	4
41	2
42	4
43	4

44	4
45	4
46	2
47	4
48	4
49	4
50	4
51	4
52	4
53	4
54	2
55	4
56	4
57	4
58	4
59	4
60	4
61	4
62	4
63	4
64	4
65	4
66	4
67	2
68	4
69	2
70	4
71	4
72	1
73	2
74	4
75	4
76	3
77	4
78	4
79	4
80	4
81	4
82	1

83	1
84	4
85	4
86	1
87	4
88	4
89	1
90	4
91	4
92	4
93	4
94	4
95	4
96	4
97	4
98	1
99	4
100	4
101	4
102	1
103	4
104	4
105	4
106	4
107	4
108	4
109	4
110	4
111	2
112	4
113	2
114	4
115	2
116	4
117	4
118	2
119	1
120	4
121	4

122	2
123	2
124	4
125	4
126	4
127	4
128	4
129	4
130	4
131	4

Table A.3: Cluster assignment for the answers to question three.

Document	Cluster
1	1
2	2
3	1
4	2
5	1
6	1
7	1
8	2
9	1
10	3
11	1
12	1
13	2
14	1
15	1
16	1
17	1
18	2
19	3
20	1
21	3
22	1
23	1
24	1

25	1
26	1
27	1
28	2
29	1
30	1
31	1
32	1
33	1
34	1
35	1
36	1
37	2
38	1
39	2
40	1
41	3
42	1
43	4
44	1
45	1
46	1
47	2
48	1
49	2
50	2
51	1
52	1
53	2
54	4
55	2
56	3
57	2
58	1
59	1
60	2
61	1
62	1
63	1

64	1
65	1
66	1
67	1
68	1
69	1
70	1
71	1
72	1
73	1
74	1
75	1
76	1
77	1
78	1
79	1
80	2
81	1
82	1
83	1
84	1
85	1
86	1
87	1
88	2
89	1
90	1
91	1
92	4
93	2
94	1
95	1
96	1
97	1
98	3
99	3
100	1
101	1
102	1

103	1
104	1
105	1
106	1
107	1
108	1
109	1
110	1
111	4
112	1
113	1
114	1
115	1
116	1
117	1
118	1
119	1
120	1
121	1
122	1
123	1
124	2
125	3
126	2
127	2
128	1
129	1
130	1
131	1

Table A.4: Cluster assignment for the answers to question four.

Document	Cluster
1	2
2	2
3	2
4	2
5	1

6	2
7	2
8	2
9	2
10	2
11	2
12	2
13	2
14	2
15	2
16	2
17	1
18	2
19	2
20	2
21	2
22	2
23	2
24	2
25	2
26	2
27	2
28	2
29	2
30	2
31	2
32	2
33	2
34	2
35	2
36	2
37	2
38	2
39	2
40	2
41	2
42	2
43	1
44	2

45	3
46	2
47	2
48	2
49	2
50	2
51	2
52	2
53	2
54	2
55	2
56	2
57	2
58	2
59	2
60	2
61	2
62	1
63	2
64	3
65	2
66	1
67	2
68	2
69	2
70	2
71	2
72	2
73	2
74	2
75	2
76	2
77	2
78	2
79	3
80	1
81	2
82	2
83	3

84	2
85	2
86	2
87	2
88	2
89	2
90	2
91	2
92	1
93	2
94	2
95	2
96	2
97	2
98	2
99	1
100	2
101	2
102	2
103	2
104	2
105	2
106	2
107	2
108	2
109	2
110	2
111	2
112	2
113	2
114	2
115	2
116	2
117	1
118	2
119	2
120	2
121	2
122	2

123	2
124	2
125	2
126	2
127	2
128	2
129	2
130	2
131	2

Table A.5: Cluster assignment for the answers to question five.

Document	Cluster
1	3
2	3
3	3
4	3
5	1
6	3
7	3
8	3
9	3
10	3
11	3
12	3
13	3
14	3
15	2
16	3
17	3
18	3
19	3
20	3
21	3
22	3
23	3
24	3
25	3

26	3
27	3
28	1
29	3
30	3
31	3
32	3
33	3
34	3
35	1
36	3
37	1
38	3
39	3
40	3
41	2
42	1
43	3
44	3
45	3
46	3
47	1
48	3
49	3
50	3
51	3
52	3
53	1
54	1
55	3
56	3
57	3
58	3
59	3
60	3
61	3
62	3
63	3
64	3

65	3
66	3
67	3
68	3
69	2
70	3
71	3
72	3
73	3
74	3
75	3
76	3
77	3
78	3
79	3
80	3
81	3
82	3
83	3
84	3
85	1
86	3
87	2
88	3
89	3
90	3
91	2
92	3
93	3
94	3
95	3
96	3
97	3
98	3
99	3
100	3
101	3
102	3
103	3

104	3
105	3
106	3
107	3
108	3
109	3
110	3
111	3
112	3
113	3
114	3
115	2
116	3
117	3
118	2
119	3
120	3
121	3
122	3
123	3
124	3
125	2
126	3
127	3
128	2
129	3
130	3
131	1

Table A.6: Cluster assignment for the answers to question six.

Document	Cluster
1	4
2	2
3	2
4	2
5	2
6	2

7	2
8	2
9	4
10	2
11	2
12	2
13	2
14	2
15	2
16	2
17	2
18	2
19	2
20	2
21	3
22	2
23	2
24	4
25	4
26	2
27	4
28	2
29	2
30	2
31	2
32	2
33	2
34	2
35	2
36	2
37	2
38	4
39	2
40	2
41	2
42	1
43	2
44	3
45	2

46	2
47	2
48	2
49	2
50	1
51	2
52	2
53	2
54	2
55	2
56	2
57	2
58	2
59	3
60	2
61	2
62	2
63	2
64	4
65	2
66	2
67	2
68	2
69	2
70	3
71	2
72	4
73	2
74	2
75	2
76	2
77	2
78	2
79	2
80	2
81	2
82	2
83	2
84	2

85	2
86	2
87	2
88	2
89	2
90	2
91	1
92	1
93	2
94	2
95	2
96	2
97	2
98	2
99	2
100	2
101	2
102	4
103	2
104	2
105	2
106	2
107	2
108	2
109	1
110	4
111	2
112	3
113	2
114	2
115	4
116	2
117	2
118	2
119	2
120	4
121	3
122	2
123	2

124	2
125	2
126	4
127	2
128	2
129	1

Figure A.1: Document clustering of answers to question one.

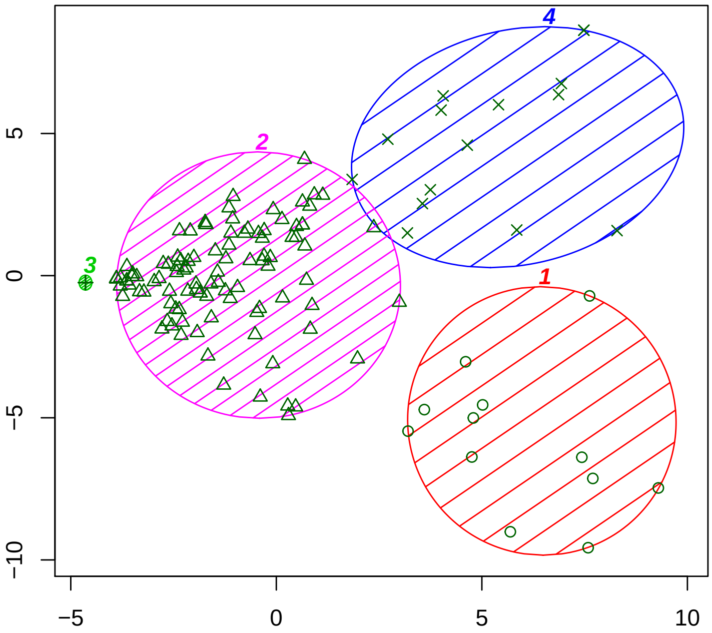


Figure A.2: Document clustering of answers to question two.

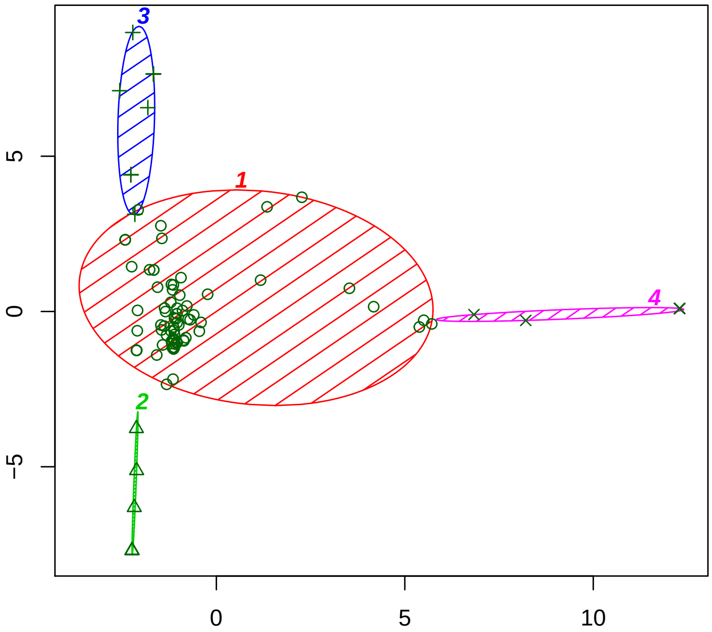


Figure A.3: Document clustering of answers to question three.

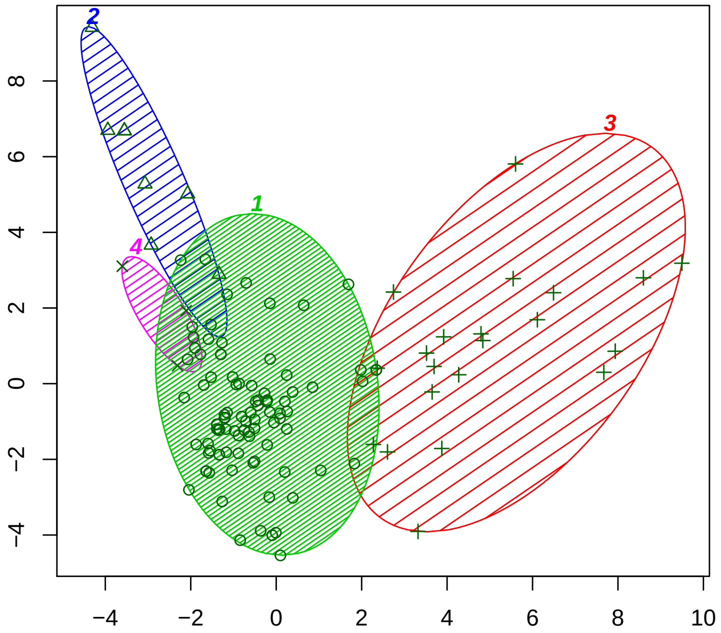


Figure A.4: Document clustering of answers to question four.

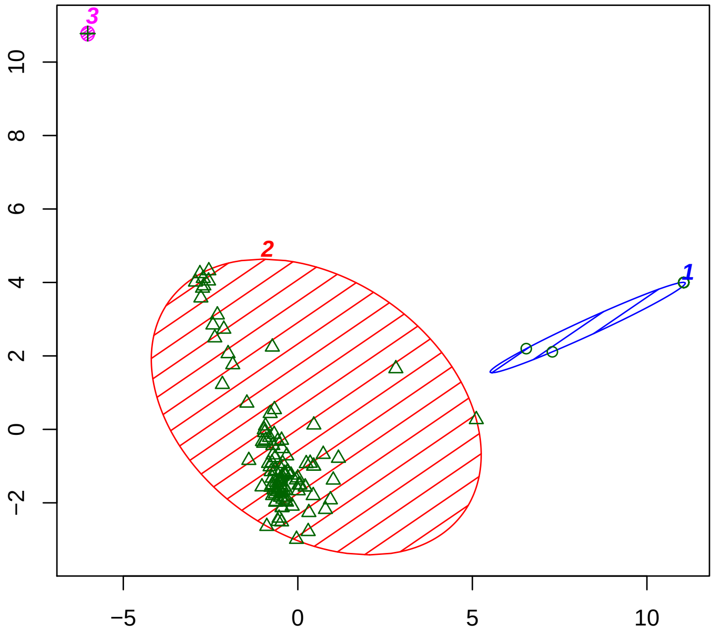


Figure A.5: Document clustering of answers to question five.

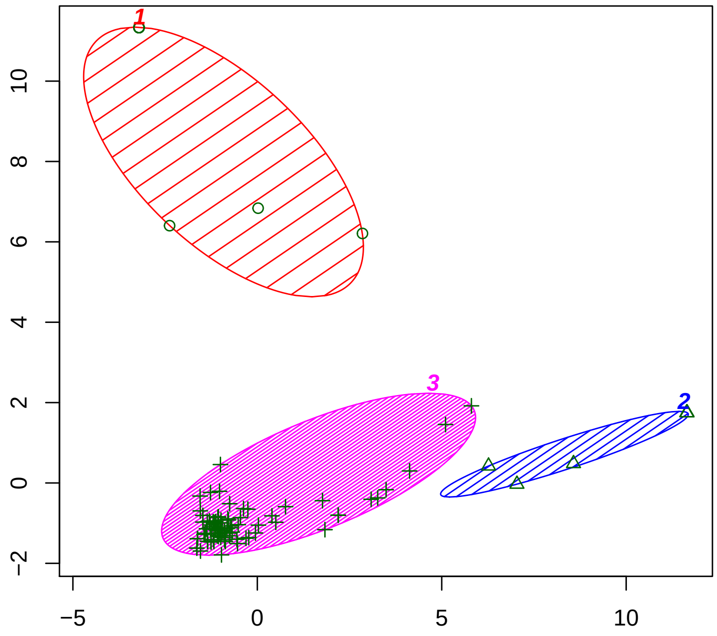
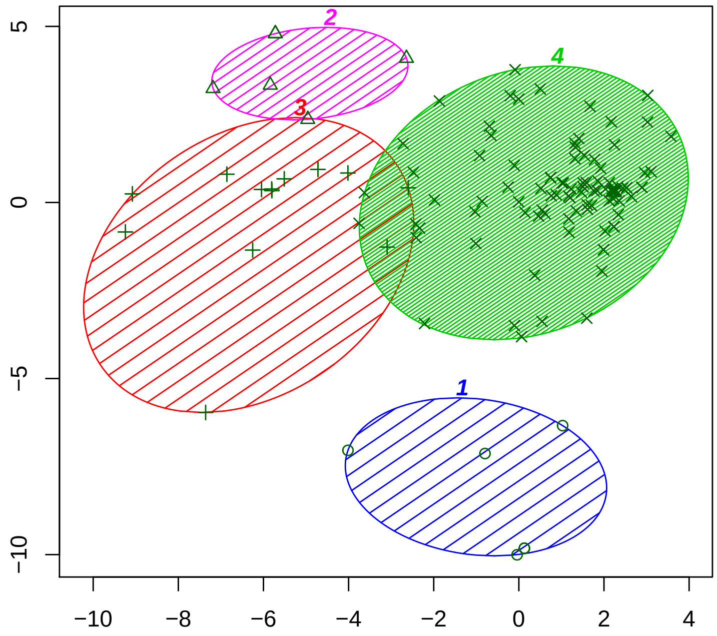


Figure A.6: Document clustering of answers to question six.



Appendix B

LDA tuning

The parameters for the topic models were chosen to be in line with the recommendation from the package *'ldatuning'*. This package compiles and displays several metrics regarding topic quality. For a full reference see `https://cran.r-project.org/web/packages/ldatuning/vignettes/topics.html`
.

The optimal number of topics should lie between the extrema of the graphics, therefore four was chosen as a suitable number of topics for all questions, in order to have betas have the same weight regarding topics between different questions.

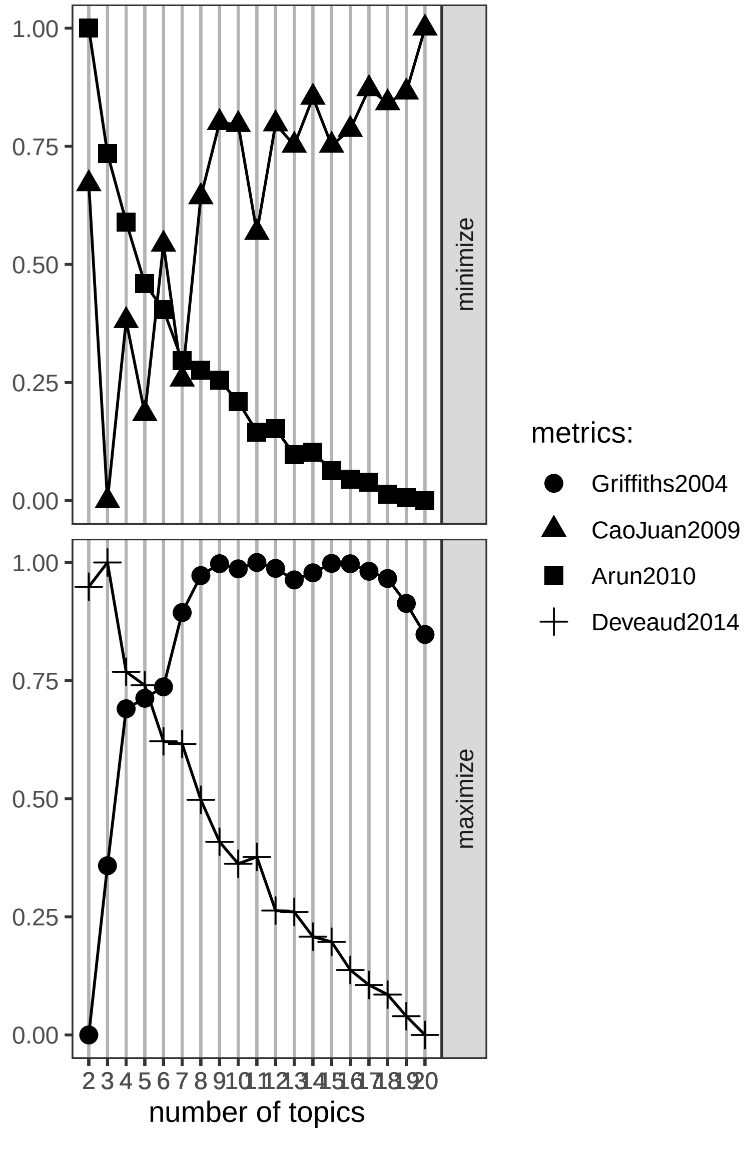


Figure B.1: Tuning for answers to question 1.

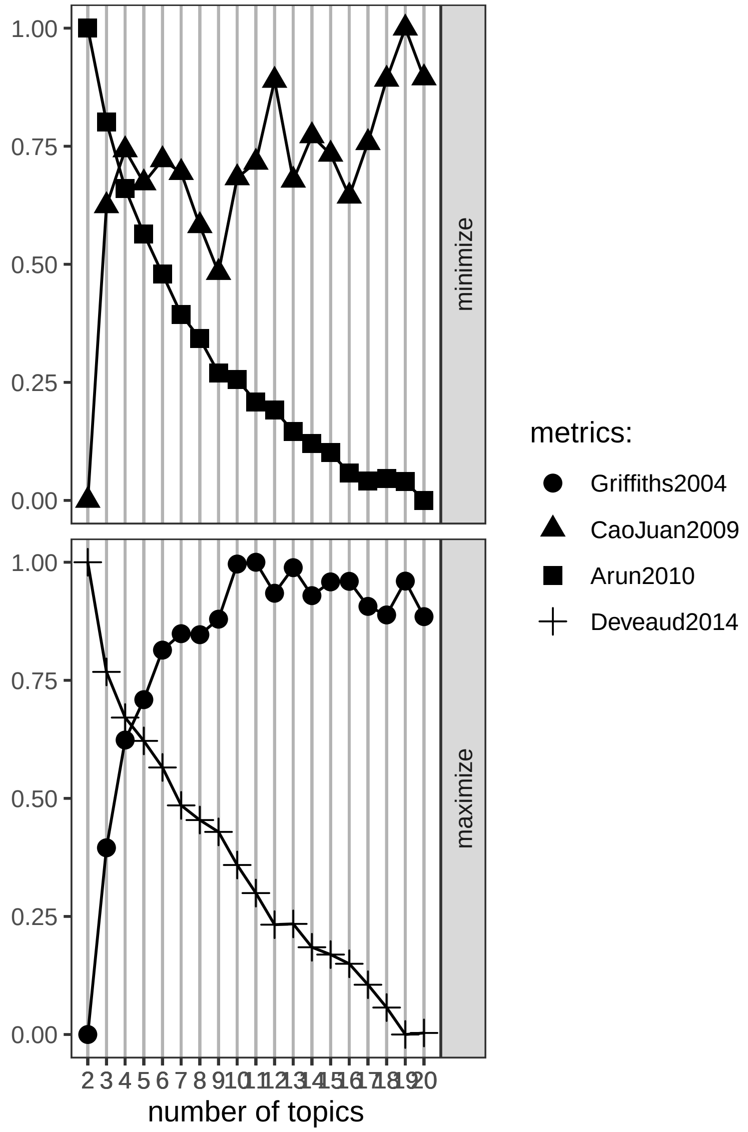


Figure B.2: Tuning for answers to question 2.

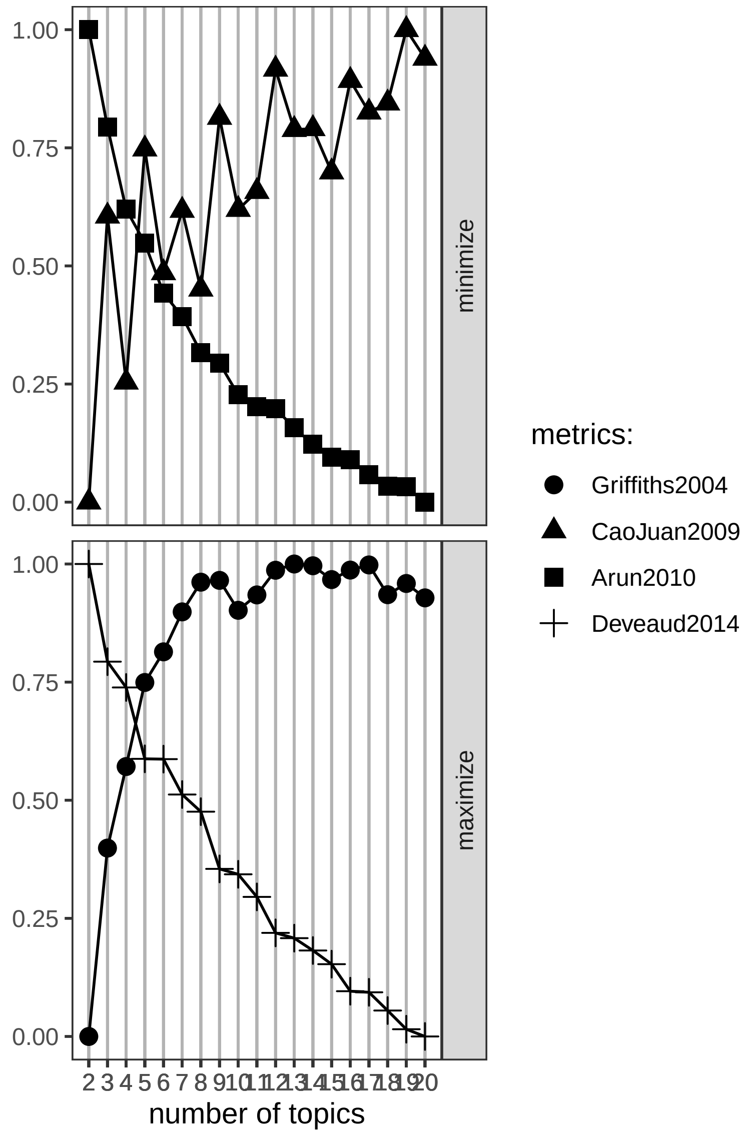


Figure B.3: Tuning for answers to question 3.

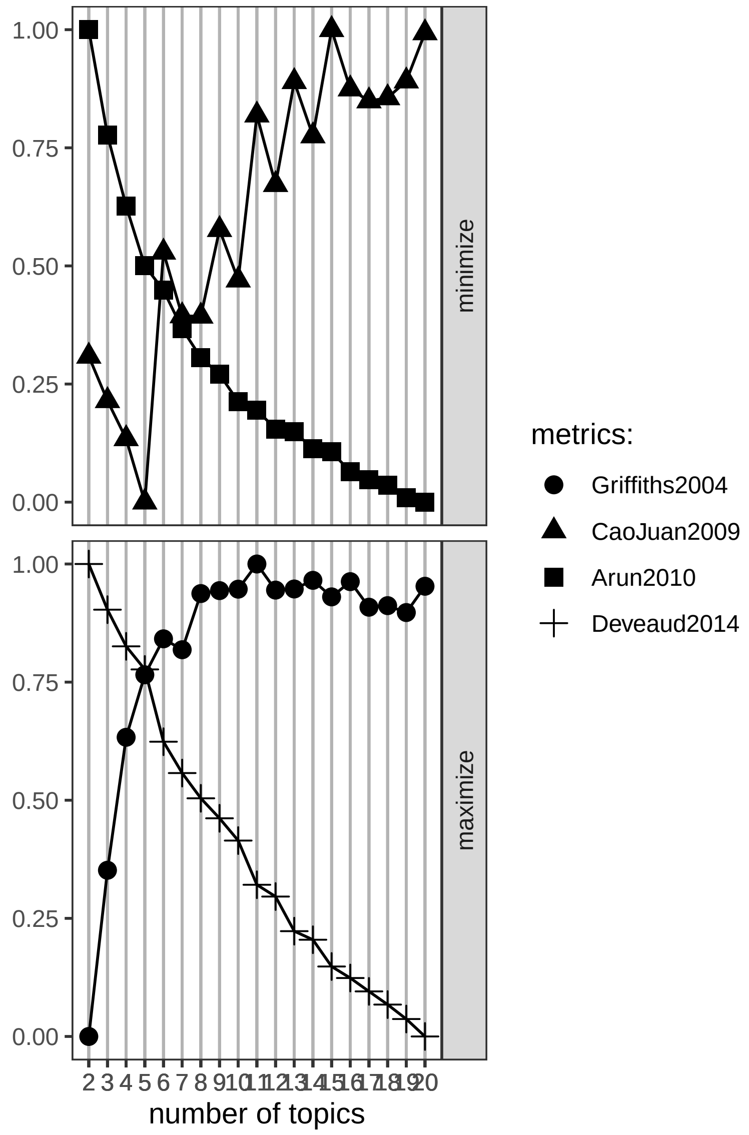


Figure B.4: Tuning for answers to question 4.

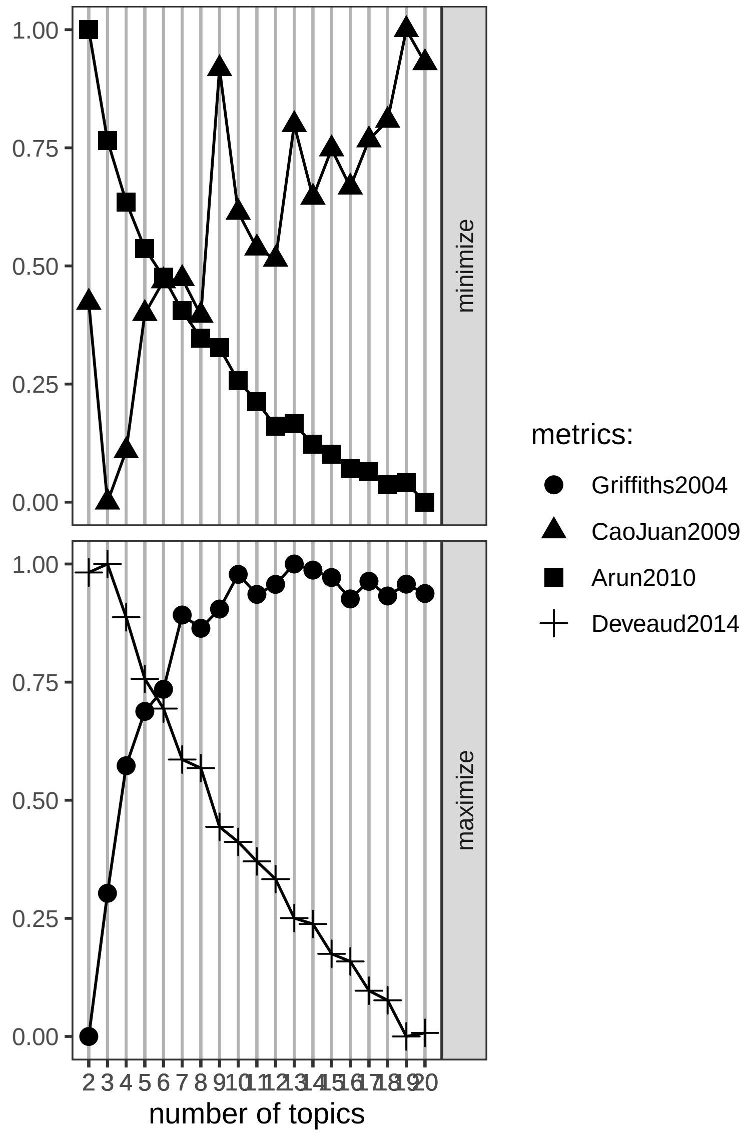


Figure B.5: Tuning for answers to question 5.

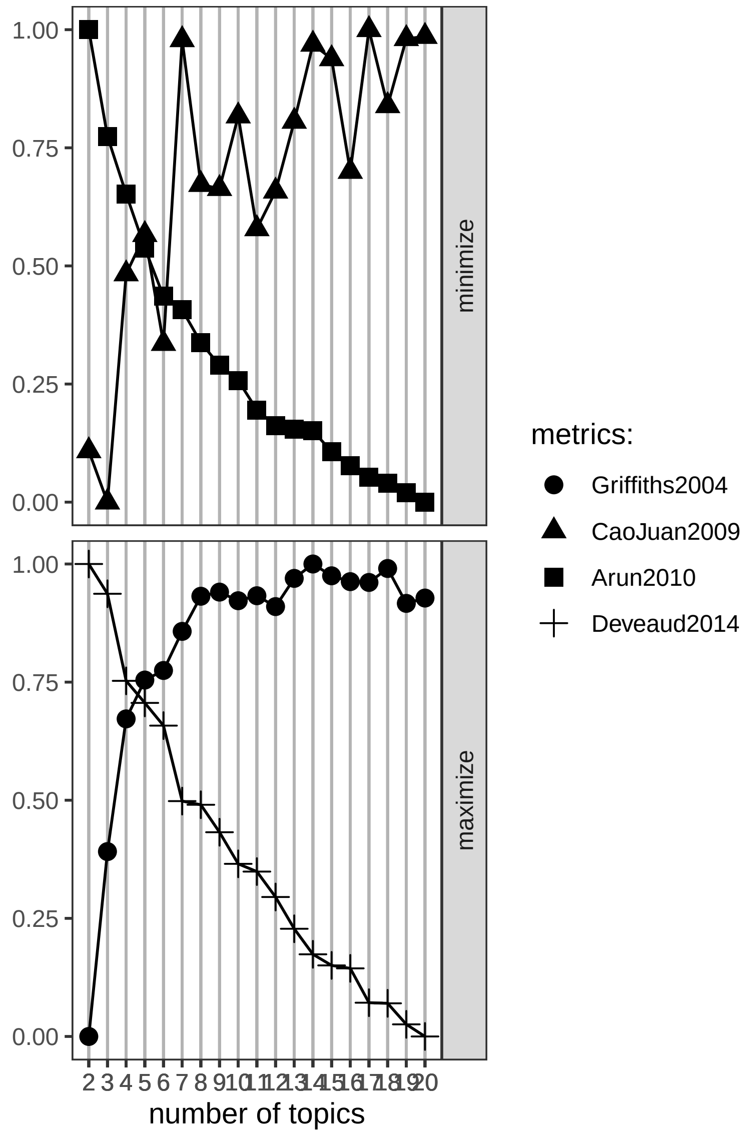
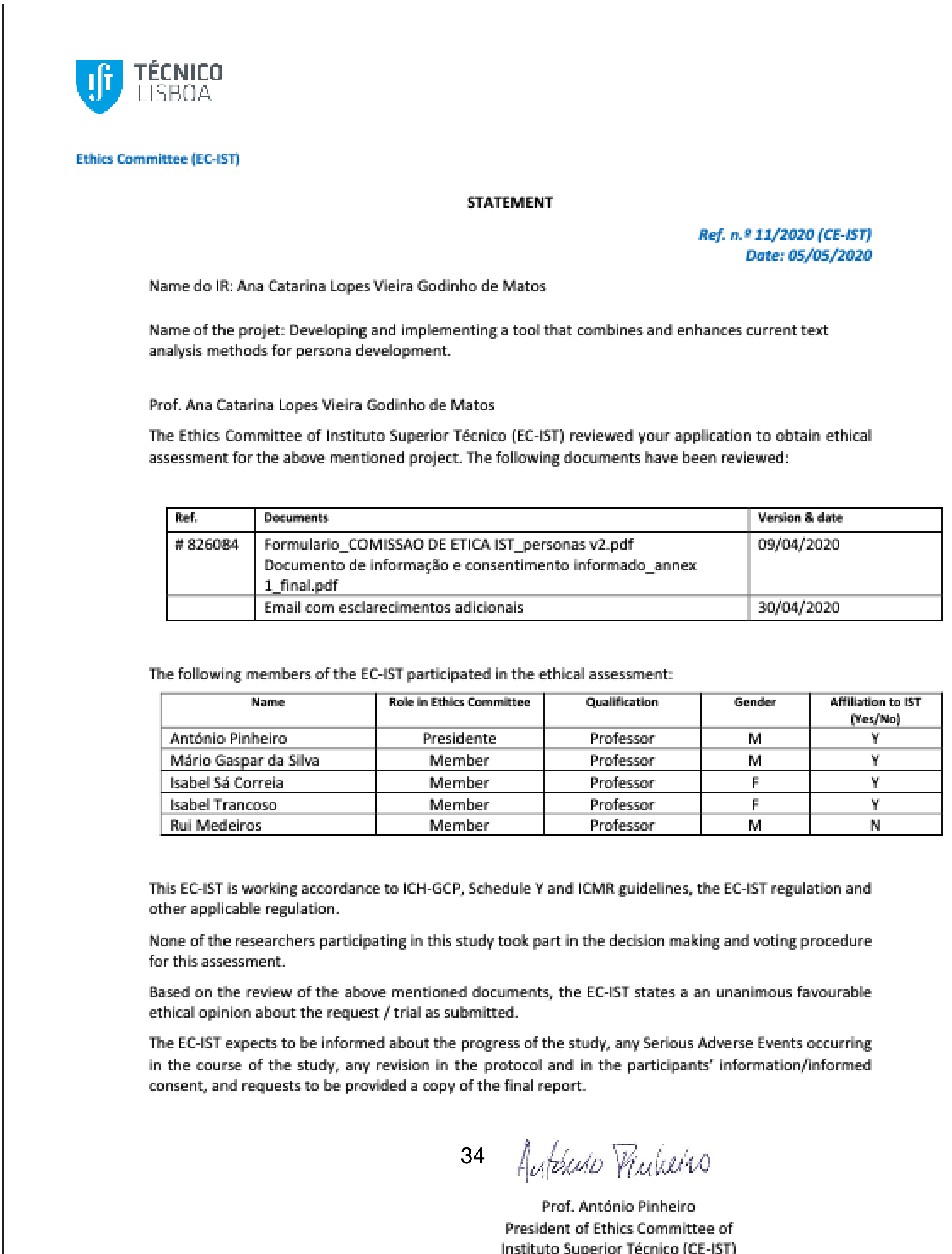


Figure B.6: Tuning for answers to question 6.

Appendix C

Ethics commission permission



TÉCNICO LISBOA

Ethics Committee (EC-IST)

STATEMENT

Ref. n.º 11/2020 (CE-IST)
Date: 05/05/2020

Name do IR: Ana Catarina Lopes Vieira Godinho de Matos

Name of the projet: Developing and implementing a tool that combines and enhances current text analysis methods for persona development.

Prof. Ana Catarina Lopes Vieira Godinho de Matos

The Ethics Committee of Instituto Superior Técnico (EC-IST) reviewed your application to obtain ethical assessment for the above mentioned project. The following documents have been reviewed:

Ref.	Documents	Version & date
# 826084	Formulario_COMISSAO DE ETICA IST_personas v2.pdf Documento de informação e consentimento informado_annex 1_final.pdf	09/04/2020
	Email com esclarecimentos adicionais	30/04/2020

The following members of the EC-IST participated in the ethical assessment:

Name	Role in Ethics Committee	Qualification	Gender	Affiliation to IST (Yes/No)
António Pinheiro	Presidente	Professor	M	Y
Mário Gaspar da Silva	Member	Professor	M	Y
Isabel Sá Correia	Member	Professor	F	Y
Isabel Trancoso	Member	Professor	F	Y
Rui Medeiros	Member	Professor	M	N

This EC-IST is working accordance to ICH-GCP, Schedule Y and ICMR guidelines, the EC-IST regulation and other applicable regulation.

None of the researchers participating in this study took part in the decision making and voting procedure for this assessment.

Based on the review of the above mentioned documents, the EC-IST states a an unanimous favourable ethical opinion about the request / trial as submitted.

The EC-IST expects to be informed about the progress of the study, any Serious Adverse Events occurring in the course of the study, any revision in the protocol and in the participants' information/informed consent, and requests to be provided a copy of the final report.

António Pinheiro

Prof. António Pinheiro
President of Ethics Committee of
Instituto Superior Técnico (CE-IST)