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

Manuel Joana de Sousa Prata
manuel.s.prata@ist.utl.pt

Instituto Superior Técnico, Lisbon, Portugal

November 2020

Abstract

In this work a methodology for persona development is created and implemented. This methodology makes use of multiple computational text analysis techniques, it is then applied to a case study in which personas are 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 the 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. There are some limitations in this present work, they can however be improved in future work, the 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

1. Introduction

In this section a brief overview of the motivation behind this work is presented. On top of this costumer personas, a concept imported from software design and marketing, are also introduced. They were introduced by Cooper in his work [3]. In this work the personas represent students instead of costumers, while keeping all the other attributes.

This work sets out to create an efficient development process for personas. 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 [19].

We develop personas from textual input using tools and techniques from the Natural Language Processing (NLP) research field. 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 [24, 26, 28].

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 [7, 28]. This allows for the creation of a persona development process based on the computational analysis of textual information addressing the issues previously mentioned.

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.

In the following sections theoretical background for this work will be presented. This section starts by introducing some elements regarding persona design followed by introducing the needed ML and NLP concepts. Afterwards, there is a section regarding the implementation of the methodology, followed by the presentation of obtained results. Finally, the personas are presented. In the closing section issues with the current work are discussed

as well as possibilities for future work.

2. Background

2.1. Persona Design

This work will focus on customer personas, which will be referred to simply as personas, from now on, 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 [25].

2.2. Persona Development Process

Customer personas are usually created by researching potential users, patterns of behaviour are discovered during research [3]. 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. 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 [8].

This approach has some limitations, namely, the fact that it is an expensive and slow methodology to apply correctly [9].

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 [32]. The methodology from [32] consists 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.

2.3. Machine Learning

Machine Learning (ML) is the study of computer algorithms that allow computer programs to automatically improve through experience [20]. 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 [21].

ML 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.

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, organise it (a clustering algorithm) [18]. Unsupervised learning techniques are able to find patterns in data automatically, which is useful for finding connections in large volumes of data [17].

In this work only unsupervised techniques were used, therefore those shall be the focus of the following paragraphs.

2.3.1 Unsupervised Learning

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 [21].

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 [12].

The cosine distance is a popular metric for textual content [27].

2.4. Natural Language Processing

Natural Language Processing 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 [13].This field is comprised by 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 [23].

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 [6].

2.4.1 Fundamental concepts of NLP

In processing text with the aid of computational methods, assumptions will inevitably be made. In the following chapters we will mention the more general ones, mentioning some properties of words in a document, namely their distribution, which obeys a power law [13].

2.4.2 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 [16].

Firstly, studying the statistical patterns of human word usage can be used to understand meaning. The study of this phenomena is called statistical linguistics [33]. This is the encompassing concept for the distributional hypothesis stated below.

2.4.3 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 [34], which will be discussed later in this text.

2.5. Text Pre-Processing

2.5.1 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 [30]. The tables take the form of one token per row.

2.5.2 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 [30]. An option is to include punctuation and other special characters. These have different uses and as such influence the choice [14].

2.5.3 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 termination of words, normally derivation affixes, such as the "-*s*" commonly added to form the plural of a word.

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 [29].

2.6. Standard NLP activities

In this section, we shall expose the techniques that 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 form the textual data available, making them suitable for the purpose of the case study.

2.6.1 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.

The heuristic 'term frequency'-'inverse document frequency', $td-idf$, is a measure of how important a word is to a document in a collection (or corpus) of documents. $td-idf$ consists of two separate statistics that are multiplied together, and in which each term captures a different concept about the statistical distribution of text [11].

The tf is calculated as

$$tf(W) = n/V \tag{1}$$

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

$$idf(W) = ln\frac{n_{documents}}{n_{documents\ containing\ w}} \tag{2}$$

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 [22].

2.6.2 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 [35].

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 [4].

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 [1].

Lexicon-based approaches use NLP and lexical resources to assign sentiment. They use mainly part-of-speech information and WordNet [1]. These approaches create a sentiment lexicon in which words are scored according to their sentiment.

2.7. Summarizing 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.

3. Implementation

3.1. Methodological Steps

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

In order to collect the data an online survey was 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 from 22/04/2020 up to 09/07/2020.

3.2. Tool

The tool has a sequential workflow until a normalized representation of the text is obtained.

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.

In the following table we can see which preprocessing steps were applied for each specific technique.

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 1: Pre-processing done per analysis task

The analysis was performed sequentially following the order of tasks present in Table 1 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.

Finally, after having analysed the data, there was a persona development step, in which personas were created according to the analysis.

3.3. Persona Generation

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

Sentiment scores will be computed for each answer. From this a table of sentiment scores was constructed and analysed.

Finally a narrative is constructed for each persona.

4. Results

4.1. Dataset Characteristics

In this work, 162 questionnaires were collected. Of these, 31 were not suitable for analysis due to various factors, this leaves us with 131 to be analysed.

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.

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 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 2: Distribution of the curricular year of each respondent.

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

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 Physics (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 3: Degree and gender distribution.

From tables 2 and 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 addressed in the closing section of this work.

4.2. Sentiment Analysis

Sentiment scores were calculated using the function *'analyzeSentiment'* from R's package *'SentimentAnalysis'*. The scores were calculated using the 'GI' dictionary[1], as it is suited for this type of textual content.

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.

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

Question	I	II	III	IV	V	VI
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 4: Fraction of (positive,negative,neutral) sentiment scores sorted by question.

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

Summing the 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.

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 [5, 2]. 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 4 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. These 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.

4.3. 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: I, II, III and VI. For questions IV and V 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 in this work as tables which present the names and sizes of each cluster, for each question. The attributed names and sizes each cluster are reported tables two through nine, in the Data Analysis section.

5. Data Analysis

5.1. Sentiment Analysis

From table 4 we can see that Neutral is the most common class in almost all answers to all questions.

Furthermore, we can take the sum of score $\frac{31,729}{131} = 0,242$[2]. Then comparing this with the sentiment score value of 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
Total	131

Table 5: 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.

5.2. Cluster Analysis

A first glance quickly gives us important information, 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. 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 6: Size of the clusters, question one answers.

[2]The sum of score is the score of questions I + II, etc, for all questions, per respondent. 31729 is the sum of sum of scores, for all respondents.

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

Table 7: Size of the clusters, question two answers.

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

Table 8: Size of the clusters, question three answers.

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

Table 9: Size of the clusters, question four answers.

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

Table 10: Size of the clusters, question five answers.

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

Table 11: Size of the clusters, question six answers.

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.

In Table 6 the third cluster is labeled NA as it contained three identical documents (same single word answer).

From table 6 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 7, relating to the answer to question 2, 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. This reveals a trend, while some respondents have negative feelings towards the institution some of them have positive ones, creating a clear avenue to separate respondents.

In the clusters from the answers to question 4 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.

6. 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.

6.1. Personas

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.

6.2. Personas

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 12: '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.

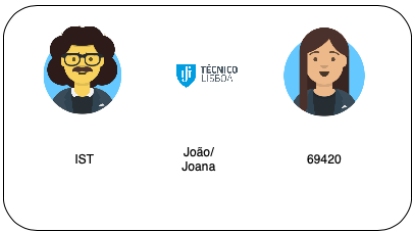

Figure 1: 'Scholar' persona student ID card.

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/Shr 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 13: 'Dominator' persona.

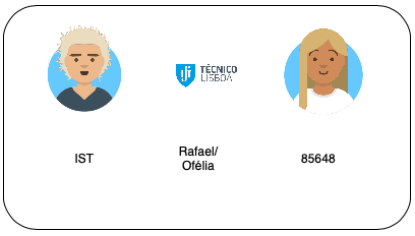

Figure 2: 'Dominator' persona student ID card.

The bio for this persona is:

He/She is a fourth year student. He/She has had excellent grades so far. He/She focuses on course-

work 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 14: 'Stable Job' persona.

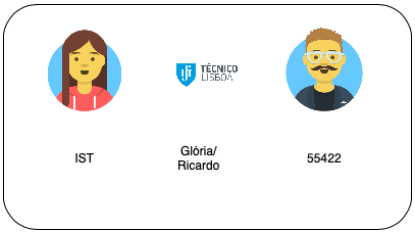

Figure 3: 'Stable Job' persona student ID card.

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 his/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, ideas for future work are presented, 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.

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.

6.3. 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.

Reading the answers given, the 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.

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.

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.

The main advantages gained over traditional 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.

6.4. 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. It should be noted, however, that these problems do not make impossible the task at hand of testing the persona design method proposed using real data.

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, 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.

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.

6.5. 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 two major techniques applied to this work were: SA and Clustering.

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.

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 [15], which makes use of BERT, a ML based technique to generate feature vectors, improving performance.

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. While this 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 [31, 10].

6.6. Closing Remarks

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

This work 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 preliminary 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.

Acknowledgements

The author would like to thank ...

References

[1] C. Anas, C. Crina, J. Damien, H. Omar, and B. Lionel. A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation. *Research report RR-LIRIS-2014-002*, 2013.

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

[3] A. Cooper. *The Inmates Are Running the Asylum*. Macmillan Publishing Co., Inc., Indianapolis, IN, USA, 1999.

[4] M. Devika, C. Sunitha, and A. Ganesh. Sentiment analysis: A comparative study on different approaches. *Procedia Computer Science*, 87:44–49, 2016.

[5] P. S. Dodds, E. M. Clark, S. Desu, M. R. Frank, A. J. Reagan, J. R. Williams, L. Mitchell, K. D. Harris, I. M. Kloumann, J. P. Bagrow, K. Megerdoomian, M. T. McMahon, B. F. Tivnan, and C. M. Danforth. Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, 112(8):2389–2394, Feb. 2015.

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

[7] D. Giorgetti, I. Prodanof, and F. Sebastiani. 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)*, pages 173–184, 2003.

[8] K. Goodwin. *Designing for the Digital Age: How to Create Human-Centred Products and Services*. Wiley, 2009.

[9] C. G. Hill, M. Haag, A. Oleson, C. Mendez, N. Marsden, A. Sarma, and M. Burnett. Gender-inclusiveness personas vs. stereotyping. In *Proceedings of the 2017 Conference on Human Factors in Computing Systems*. ACM Press, 2017.

[10] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-efficient transfer learning for nlp, 2019.

[11] D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition, 2000.

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

[13] B. E. King and K. Reinold. Natural language processing. In *Finding the Concept, Not Just the Word*, pages 67–78. Elsevier, 2008.

[14] A. Kumar. *Mastering Text Mining with R*. Packt Publishing, dec 2016.

[15] Y. Li, J. Cai, and J. Wang. A text document clustering method based on weighted bert model. In *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, volume 1, pages 1426–1430, 2020.

[16] B. F. Ljungberg. Dimensionality reduction for bag-of-words models : Pca vs lsa. 2017.

[17] Z. Madhoushi, A. R. Hamdan, and S. Zainudin. Sentiment analysis techniques in recent works. In *2015 Science and Information Conference (SAI)*. IEEE, July 2015.

[18] S. Marsland. *Machine Learning: An Algorithmic Perspective, Second Edition*. Chapman & Hall/CRC, 2nd edition, 2014.

[19] J. J. McGinn and N. Kotamraju. Data-driven persona development. In *Proceeding of the twenty-sixth annual conference on Human factors in computing systems - '08*. ACM Press, 2008.

[20] T. M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

[21] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[22] E. Nguyen. Text mining and network analysis of digital libraries in r. In *Data Mining Applications with R*, pages 95–115. Elsevier, 2014.

[23] I. Nitin and D. Fred. *Handbook of Natural Language Processing*. 2010.

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

[25] J. Pruitt and T. Adlin. *The Persona Lifecycle: Keeping People in Mind Throughout Product Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

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

[27] L. Sahu and B. R. Mohan. 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, Dec. 2014.

[28] M. Schmidt. 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, July 2010.

[29] A. Schofield, M. Magnusson, and D. Mimno. 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, 2017.

[30] J. Silge. *Text Mining with R: A Tidy Approach*. O'Reilly Media, jul 2017.

[31] A. C. Stickland and I. Murray. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning, 2019.

[32] V. Thoma and B. Williams. Developing and validating personas in e-commerce: A heuristic approach. pages 524–527, 08 2009.

[33] P. D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, Feb. 2010.

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

[35] L. Zhang and B. Liu. *Sentiment Analysis and Opinion Mining*. Springer US, 2016.