Master Thesis in Electronic Government
University of Fribourg, Department of Informatics

Candidate’s Profile Parameters Definition for VAAs-Matching Algorithms.

Author: José Alberto Mancera Andrade
Student Number: 10-801-207
Research Group: Information Systems DIUF
Examiner: Prof. Dr. Andreas Meier
Supervisor: Dr. Luis Terán
# Table of contents

1. Introduction .................................................................................................................. 3
   1.1 Problem statement .................................................................................................. 3
   1.2 Research Objectives and Methodology ................................................................. 3
      1.2.1 Research Questions ...................................................................................... 3
      1.2.2 Research Methodology .............................................................................. 4
   1.3 Timetable ............................................................................................................... 4
   1.4 Addressees ............................................................................................................ 4
2. Voting Advice Applications VAAs .............................................................................. 5
   2.1 Basic Definition .................................................................................................... 5
   2.2 VAAs Influence and Importance .......................................................................... 5
   2.3 VAAs Impartiality Challenges .............................................................................. 6
3. Candidate Profile ......................................................................................................... 8
   3.1 General Considerations and Assumptions .......................................................... 8
   3.2 Candidate Profile Parameter Categories ............................................................. 8
   3.3 Candidate Profile Model ..................................................................................... 9
   3.4 Parameters Weight Categorization ...................................................................... 9
   3.5 Candidate Profile Vector ................................................................................... 12
      3.5.1 Static Candidate’s Profile Vector ............................................................... 13
      3.5.2 Dynamic Candidate’s Profile Vector ........................................................ 14
4. Dataset Selection ......................................................................................................... 15
   4.1 Datasets and Selection Criteria .......................................................................... 15
   4.2 Dataset Structure ................................................................................................. 16
      4.2.1 Ecuador Presidential 2013 .......................................................................... 17
      4.2.2 Peru Presidential 2011 ............................................................................... 18
   4.3 Dataset Pre-Processing ......................................................................................... 19
      4.3.1 Elections in Peru 2011 Candidates and Parties ........................................... 19
      4.3.2 Elections in Ecuador 2013 Candidates and Parties .................................. 20
4.4 Dataset Processing .............................................................................................................. 21

5. Sentiment Analysis ............................................................................................................. 23
  5.1 Sentiment Analysis Elements ............................................................................................ 23
  5.2. Mapping Tweets to a Vector ............................................................................................ 25
  5.3 Sentiment Spanish Dictionary ........................................................................................... 26

6. Voting Advice Application Algorithms ................................................................................... 28
  6.1 Euclidean distance ............................................................................................................. 28
  6.2 Fuzzy C-means Algorithm ................................................................................................. 29
  6.3 Principal Component Analysis (PCA) ................................................................................ 31

7. Analysis and Evaluation ......................................................................................................... 33
  7.1 Sentiment Analysis ............................................................................................................ 34
    7.1.1 Statistics of Twitter Accounts ....................................................................................... 34
    7.1.2 Constructing the Dynamic Vector ................................................................................ 40
  7.2 Candidates Vectors ............................................................................................................ 42
    7.2.1 Ecuador .......................................................................................................................... 42
    7.2.2 Peru ............................................................................................................................... 43
  7.3 Euclidian Algorithm ........................................................................................................... 45
    7.3.1 Ecuador .......................................................................................................................... 46
    7.3.2 Peru ............................................................................................................................... 48
  7.4 Dataset and Candidate Vector Integration .......................................................................... 50
    7.4.1 Ecuador .......................................................................................................................... 50
    7.4.2 Peru ............................................................................................................................... 52
  7.5 Fuzzy Clustering Algorithm .............................................................................................. 54
    7.5.1 Ecuador .......................................................................................................................... 54
    7.5.2 Peru ............................................................................................................................... 56
  7.6 Degrees of Membership .................................................................................................... 59
    7.6.1 Ecuador .......................................................................................................................... 59
    7.6.2 Peru ............................................................................................................................... 61
8. Recommendations ............................................................................................... 63
9. Conclusions ....................................................................................................... 65
10. Future Work ..................................................................................................... 66
11. References ....................................................................................................... 67
12. Appendix .......................................................................................................... 68
12.1 Word Count Words Program .......................................................................... 68
12.2 JSON To CVS Converter with UTF-8 Format .............................................. 69
12.3 Collect all the User Twitter Information From Users .................................. 70
12.4 Tweets Capture for a User Account (3200 Tweets) ..................................... 71
12.5 Twitter Stream Listener For Many Accounts ............................................. 72
12.6 Sentiment Analysis Dictionary Comparison (Code Section) ..................... 73
# Table of Figures

Figure 1. Candidate Profile ................................................................. 9  
Figure 2. Main social topic categories ................................................ 10  
Figure 3. Candidate’s Profile Vector Construction ................................ 12  
Figure 4. Dynamic Candidate’s Vector ................................................ 14  
Figure 5. Dataset Criteria Comparison ............................................... 16  
Figure 6. Dataset Structure .............................................................. 16  
Figure 7. Ecuador Social Issue Questions ........................................... 18  
Figure 8. Peru Social Issue Questions ................................................ 19  
Figure 9. Elections in Peru 2011 Candidates and Parties ....................... 20  
Figure 10. Elections in Ecuador 2013 Candidates and Parties ............... 20  
Figure 11. Dataset Cleaning Process .................................................. 21  
Figure 12. Dataset Processing for Peru and Ecuador ........................... 22  
Figure 13. Sentiment Analysis Architecture ....................................... 24  
Figure 14. Tweet’s Capture, Word Count System and Mapped Vector ...... 26  
Figure 15. Corresponding weights between dictionaries ....................... 27  
Figure 16. Scenario A without Sentiment Analysis ............................... 33  
Figure 17. Scenario B with Sentiment Analysis ................................... 33  
Figure 18. Followers of Ecuador Candidates ..................................... 35  
Figure 19. Number of Tweet Status for Ecuador Candidates ................. 35  
Figure 20. Followers of Peru Candidates ........................................... 36  
Figure 21. Number of Tweet Status for Peru Candidates ....................... 36  
Figure 22. Peru’s Candidate Tweets - Top Words ................................ 37  
Figure 23. Ecuador’s Candidate Tweets - Top Words ............................ 38  
Figure 24. Twitter References ........................................................... 39  
Figure 25. Ecuador Dynamic Candidates’ Vectors ............................... 41  
Figure 26. Peru Dynamic Candidates’ Vectors .................................... 42  
Figure 27. Ecuador Candidates’ Profile Vectors ................................... 43  
Figure 28. Peru Candidates’ Profile Vectors ....................................... 44  
Figure 29. Euclidian Distance Process ................................................. 45  
Figure 30. Vote Intention (Sup Question 3) .......................................... 46  
Figure 31. Euclidian Distance – Ecuador ............................................ 46  
Figure 32. Euclidian Distance (Sentiment Analysis) - Ecuador .............. 47  
Figure 33. Vote Intention (Sup Question 5) .......................................... 48
Figure 34. Euclidian Distance – Peru ................................................................. 48
Figure 35. Euclidean Distance (Sentiment Analysis) – Peru............................... 49
Figure 36. Initial Candidates' Centers from Experts (No Sentiment Analysis) Ecuador .................................................................................................................................................................................. 50
Figure 37. Initial Candidates' Centers from Experts (Sentiment Analysis) Ecuador. 51
Figure 38. Initial Candidates' Centers (No Sentiment Analysis) Peru ..................... 52
Figure 39. Initial Candidates' Centers (Sentiment Analysis) Peru .......................... 53
Figure 40. Fuzzy Clustering Algorithm (No-Sentiment Analysis) Ecuador ............ 54
Figure 41. Fuzzy Clustering Algorithm Voters' Clusters Ecuador ............................. 55
Figure 42. Fuzzy Clustering Algorithm (Sentiment Analysis) Ecuador ................. 55
Figure 43. Fuzzy Clustering Algorithm (No-Sentiment Analysis) Peru ................ 56
Figure 44. Fuzzy Clustering Algorithm Voters' Clusters (No-Sentiment Analysis) Peru .................................................................................................................................................................................. 57
Figure 45. Fuzzy Clustering Algorithm (Sentiment Analysis) Peru ....................... 58
Figure 46. Fuzzy Clustering Algorithm Voters' Clusters (Sentiment Analysis) Peru . 58
Figure 47. Degrees of Membership (No-Sentiment Analysis) Ecuador ............... 59
Figure 48. Degrees of Membership (Sentiment Analysis) Ecuador ...................... 60
Figure 49. Degrees of Membership (No-Sentiment Analysis) Peru ...................... 61
Figure 50. Degrees of Membership (Sentiment Analysis) Peru ............................ 62
Figure 51. Real Time Sentiment Analysis Architecture........................................ 66
1. Introduction

1.1 Problem statement

Overview

In the context of elections the Internet opens new and promising possibilities for parties and candidates wanting to introduce themselves and their political strategy, to organize the election campaign, to gather funds, to mobilize support and to enter into a direct dialogue with the electorate. Nowadays one element of particular importance is so-called Voting Advice Applications (VAAs) that have proliferated all over the world. VAAs are web applications that offer help in deciding how to vote in elections by comparing the preferences of parties or candidates with respect different political issues and preferences of specific voters. VAAs can determine those parties or candidates who are politically close.

VAAs can make smaller the gap between political candidates and voters, and make easier the task of political parties in order to avoid losing the visibility of the voters’ preferences, so that their strategies can be more effective during a political campaign.

The goal of this thesis is to develop a candidate profile based on different parameters such as: perspective of the voters, social networks and candidate or experts opinions. Understanding the elements that compose a candidate profile will help future candidates to develop a more successful political campaign or model to develop a better public image with respect the perception from the voters.

1.2 Research Objectives and Methodology

1.2.1 Research Questions

1. What are the parameters and their weight in a candidate’s profile?
2. Does sentiment analysis has an impact in a candidate’s profile by comparing two scenarios with a fuzzy clustering algorithm (with and without sentiment analysis)?
3. What would be the design of a generic candidate profile that relies also in sentiment analysis?
4. What are the cases and scenarios to test the proposed candidate profile?
5. What will be potential future recommendations or improvements form the candidate’s profile model?
1.2.2 Research Methodology

In a first step, selected textbooks, previous research papers and similar cases will be taken into account, in order to get an overview of the theoretical framework. The second step of this thesis will consider the proposition of a candidate’ profile structure and a recommender system model.

1.3 Timetable

<table>
<thead>
<tr>
<th>Date</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.02.2016</td>
<td>Submission of the proposal</td>
</tr>
<tr>
<td>22.02.2016</td>
<td>Start of the Master Thesis</td>
</tr>
<tr>
<td>March 2016 – Mid-May 2016</td>
<td>Revisions, writing report, research, continuous meetings with supervisors.</td>
</tr>
<tr>
<td>12.08.2016</td>
<td>Submission of the thesis report</td>
</tr>
<tr>
<td>17.08.2016</td>
<td>Thesis defense</td>
</tr>
<tr>
<td>22.08.2016</td>
<td>Delivery thesis report at the decanat</td>
</tr>
</tbody>
</table>

1.4 Addressees

The target audience of thesis is primarily students in the fields of computer science, marketing, political sciences and professionals who are involved in the field of Voting Advice Applications. The results of this thesis document should provide the parties mentioned above not only valuable knowledge in order to better understand, analyze and improve the presented candidate profile model, but also a better understanding of the consequences of the candidate profile on voters and parties.
2. Voting Advice Applications VAAs

In order to gain a better understanding of the findings presented in this master thesis, it is important to understand the VAAs concept before moving into the analysis and outcome of this work. The intention of this section is to provide an overview of VAAs and understand why they are important in the framework of political events.

2.1 Basic Definition

Voting Advice Applications (VAAs) are web-based applications, which provide information about parties or candidates running in elections. They aim to help voters to find out which party or candidate, they should vote for. Traditionally, VAAs match the voters' political preferences to those of parties and candidates running in elections in favor of provide a recommendation to the voter.

In other perspective, VAAs aim to help voters to figure out which parties or candidates share their political preferences in a cognitively easy and accessible way. Nevertheless, this analysis can be also an advantage not only for a voter who gets a recommendation, but also to a party or candidate who is interested in the elements that are needed in order to gain more registered voters.

2.2. VAAs Influence and Importance

VAAs are increasingly used in election campaigns worldwide, thus their design and methodology deserve closer attention, especially because the outputs they produce seem to affect those who use them.

In Netherlands, about 10% of users reported that they adjusted their electoral decision because of the voting recommendation they received from the application (Kleinnijenhuis & van Hoof, 2008). In Germany, 6% of participants indicated that they changed their vote choice to another party (Marschall, 2005), while 3% did so in Finland (Mykkänen & Moring, 2006). In Switzerland, a substantive amount of participants stated that the voting recommendation affected their voting behavior. Around 70% of survey participants indicated that the voting recommendation influenced their decisions on the ballot, with most of them voting for different candidates because of the ranking list that they received from the VAA. A smaller but still substantial part of users indicated that they changed their party choice as a consequence of the VAA output (Ladner et al., 2010; Ladner & Fivaz, 2012; Ladner, Fivaz, & Pianzola, 2010, Ladner, Fivaz, & Pianzola, 2012).
Although more studies in this field have to be done, all these previous studies reflect evidence that VAAs have a strong influence to persuade changes in voters decisions.

### 2.3. VAAs Impartiality Challenges

VAAs play a main role in many political events, unfortunately due to its importance it is possible to modify their recommendations (biased them) in favor of personal or political interests.

VAA effects have exclusively relied on observational studies, mostly employing user surveys to report about the impacts of the web application on voters. Such procedures are problematic because most of the samples from which results were derived consisted of self-selected study participants.

Self-selection is a problematic feature of survey research because it can severely bias the results obtained from it. If people who are enthusiastic about introducing information in the VAAs are more likely to participate in surveys on them, the resulting sample easily consists of those VAA users who report stronger effects of the application on their decisions compared to those who have not participated in the sample. Hence, we conclude strong and positive effects of VAAs on voters from a sample of strongly influenced users (Pianzola & Ladner, 2011). Contrary to the electorate considered as a whole, we currently do not know the entire population of VAA users, making it impossible to draw a random sample from the entire population of VAA users.

Without such a randomly selected sample, general inferences about VAA users from self-selected survey samples cannot be made.

In practice, in order to reduce the bias created by self-selection, the VAAs add more details about the participants in the sense to know if they belong to a political party or their level of political participation. As a consequence with these information the VAAs can classify the users in order to balance its recommendations based on user groups and establish recommendation differences among citizens, political activists and expert politicians who might try to alter the recommendation algorithm based on their answers.

Unfortunately, despite all the different techniques and best practices, this kind of control is in many cases very hard to control and most of the time impossible to achieve.
Although different studies and practical cases reviewed as documentation of this thesis have recognized that VAAs impartiality is not easy to overcome, there is an open opportunity to propose new strategies of models in the design of VAAs to reduce this bias effect in the recommendations.

One of the goals of this thesis is to decrease in part this bias effect by creating a generic candidate profile that can be harder to be biased as it will be introduced as part of the research work in further sections.
3. Candidate Profile

Problems of honesty from users, extreme candidate supporters or discrediting campaign actions are some among many situations that are difficult to control and normally outside of the scope of the design of VAAs.

After an extensive revision among different elections datasets, seminar and research papers about VAAs. This study considers that the pillar of the VAAs design should be based on a resistant or resilient candidate profile model that can tolerate answers or user manipulations in order to represent the most accurate goals, ideas or mentality of a candidate or political party.

The definition of a candidate profile model is crucial because the image of a candidate during elections, reflects the political party ideals and goals as a whole. This chapter has the aim to define the different elements that can define a candidate profile and explain the criteria behind all these characteristics.

3.1 General Considerations and Assumptions

Although surveys offer good insiders of the voters’ perceptions or preferences for a candidate, they are only valid for a short period of time and new surveys are needed to be carried on again because people tend to change opinion based on events in the society related to a candidate, such as a corruption case, conflict of interests or positive events like supporting humanitarian causes.

Nowadays the information of such events is distributed not only by traditional media (radio and TV) but also by social networks such as Twitter, Facebook, and Instagram etc. For the purpose of this study, the social network activity is considered to be a faster communication channel than traditional media, in addition surveys are considered static and valid for short period of time till there is new public information from the candidate or party.

3.2 Candidate Profile Parameter Categories

Based on all the elements and the context in which a candidate can have a public figure. This study considers a difference and define two main parameter categories:

- **Static Parameters**: All the elements that provide information that can be valid for a short period of time such as questionnaires, surveys, popularity ranking, expert opinions, and candidate’s own perception.
- **Dynamic Parameters**: All the elements that have an impact in the candidate profile at any time because of the online activity, such as twitter, Instagram, Facebook, online news blogs, YouTube etc.

The combination of both categories is a good basis for a candidate profile template that can represent all the factors involved in the candidate political environment.

### 3.3 Candidate Profile Model

The Figure 1 illustrates the parameters involved in the candidate profile such as voters, experts, candidates and social media and their classification in the profile.

![Candidate Profile Diagram]

**Figure 1. Candidate Profile**

### 3.4 Parameters Weight Categorization

As it is illustrated by the candidate model, each of the parameter categories affects one or many of different social topics that society is concerned. Based on the elections datasets considered in this work, it is possible to consider six main social topic categories or political issues. (Figure 2)
The context of these political issues and their meaning according this document are understood as follows:

- **Economy**: It refers to the perception in which the candidate or party ideas are considered by the population with respect of the national economy and policies.
- **International Policy**: It refers to the government effort to gain prestige and collaborate in an international context based on the opinions of the citizens.
- **National policy**: It considers the behavior and the perceived performance of the government and their policies in general.
- **Security**: It indicates the perception of the society related to protect the human rights, actions that reduce the levels of crime or protection against external menaces.
- **Welfare State**: It specifies the level of satisfaction that the population has in the government, this includes for example to know the way that the politicians spend the resources in the society or providing more education or more health services etc.
- **Society**: It reflects the position from the candidate or the party with respect religion and society values. For instance the government position to legalize drugs, abortion, same sex marriage etc.
It is also necessary to consider that these political issues can change their priority according to the situation in the country. For instance a deep economic recession will give more importance to the position of the government in terms of economy than society topics such as same sex marriage.

Based on these changes, some of these political issues could be more relevant than others in a certain period of time. Thus a definition of a political issue weight is needed in the profile model.
3.5 Candidate Profile Vector

As it was presented in previous sections, the candidate model has two main parameter categories (static and dynamic). In addition each social category will have a weight assigned in the final candidate’s vector by an importance vector. The Figure 3 shows the construction of a candidate profile vector by correlating all the political issues.

Figure 3. Candidate's Profile Vector Construction
In the figure 3, the political issues are mapped into three vectors of 30 elements each or dimensions. The size of the vectors is based on the original surveys provided by the datasets and each dimension is represented by a question. In addition, in some cases either the expert or candidate vector or both are available, thus for the static part, it suffices that one of them is considered.

This situation will appear in further sections, where one dataset contains only an expert vector (Ecuador) and the other only a candidate vector (Peru). Ideally considering both vectors is the best suggestion, but it can be hard to get both for a particular dataset.

Regarding the expert vector, it represents the opinion of journalists and independent politician analysts about political issues. In contrast, the static candidates’ vector represents the opinion of the candidates themselves with the respect political issues.

The final step in figure 3, is to create the candidate’s profile, so the average between the static and dynamic part is computed. Then the importance vector, considers different weights for each political issue. Each weight has a scale from 1 to 0 where 1 means important and 0 no relevant. As a first setup in our model analysis, the importance vector has a default setup of ones, thus all the social topics are considered equally important.

In the subsequent sections, more details are provided in order to understand the construction and uses of this vector in our analysis.

### 3.5.1 Static Candidate’s Profile Vector

As it was mentioned briefly before, the Figure 3 shows a static candidate vector, originally this vector is provided directly from the candidate or a political party based on each of the questions that are related to the social topic categories. Nevertheless, by analysis experience and the results shown in the seminar work (Jose A. Mancera, Bossard), this vector is the most biased among all the different elements in the model because the political interests play an important role. Therefore the candidate will try to answer in a way to get the maximum opportunity to earn popularity, regardless if his answers represent the real ideals or thoughts of the candidate. For this reason, the static candidate profile is not considered alone in our analysis.
3.5.2 Dynamic Candidate's Profile Vector

Furthermore, the dynamic candidate vector is not automatically given by a political party or candidate. It requires an extra pre-processing from the perspective of sentiment analysis. The Figure 4 shows the steps to obtain the dynamic part.

The preliminary dynamic candidate’s vector is the result of a sentiment analysis in a vector of 30 dimensions. The sentiment analysis can be applied not only to one social network but also many of them. For the purposes of this master thesis, the social network under this analysis is Twitter.

The details about the sentiment analysis, tools, scenarios, scale and methodology used to obtain the dynamic vector are properly discussed in chapter 5. For the moment is important to have the general overview and understand the structure of the proposed candidate’s vector model because this represent just one element that will play in the evaluation phase in chapter 7.
4. Dataset Selection

A profile candidate model without real data to test is not a complete analysis, hence the dataset selection and processing is another relevant element in this master thesis. Based on the information contained in the dataset, it will be possible to obtain a balanced sample of users that represent the majority of the voters for each candidate.

Dataset data extraction requires some pre-manipulation and data preparation. The following chapter describes the process in detail and also the selection criteria among four possible dataset options. In addition, the group of rules that will allow cleaning and preprocessing the selected datasets.

4.1 Datasets and Selection Criteria

The information used in this thesis was provided by preferencematcher.org, which contains several datasets of different political events. Given the nature of different datasets, this master thesis focuses on the ones in Spanish. The main reason behind this criterion is that the author and supervisor of this work have Spanish as mother tongue and it is possible to grasp in more details, understand the cultural perspective and context of the information.

The five datasets candidates to be analyzed are listed as follows:

- Spain Parliamentary 2015
- Ecuador Presidential 2013
- Catalunya Regional 2012
- Peru Presidential 2011

The main elements to take in consideration of the datasets information are:

- Dataset (CSV)
- Dataset codebook
- Candidate vector
- Expert vector
- Candidates Twitter activity
The Figure 5 presents a resume of the datasets criteria. The Spain parliamentary and Catalunya datasets present some problems in terms of completeness, both datasets do not have a candidate nor expert vectors.

<table>
<thead>
<tr>
<th>Political Event</th>
<th>Dataset (CSV)</th>
<th>Dataset (Codebook)</th>
<th>Candidate Vector</th>
<th>Expert Vector</th>
<th>Twitter Activity for Candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spain Parliamentary 2015</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Active</td>
</tr>
<tr>
<td>Ecuador Presidential 2013</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>x</td>
<td>Active with 3 passive candidates</td>
</tr>
<tr>
<td>Catalunya Regional 2012</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>Active with 4 passive candidates</td>
</tr>
<tr>
<td>Peru Presidential 2011</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>Active with 1 passive candidate</td>
</tr>
</tbody>
</table>

*Figure 5. Dataset Criteria Comparison*

Based on the information of the Figure 5, the Peru and Ecuador presidential elections remain as the more complete datasets and are ideal in order to perform some analysis and comparison.

### 4.2 Dataset Structure

The datasets contain several categories and answers provided by the voters, nevertheless in order to clean the dataset, it is necessary to identify the information that is needed, the Figure 6 illustrates the structure of the selected datasets.

The datasets are structured in two big blocks; the superior questions block (age, educational level, vote intention, favorite political party etc.) and the social topic questions which are presented in the sections 4.2.1 and 4.2.2. For analysis purposes and comparisons, the supreme question “vote intention” is considered only from the superior questions block. The rest of the supreme questions are ignored for this study but they may become relevant in the case that more granularities in the results are needed such as the perception in different educational levels for instance.

The second reason to not consider a deeper level of granularity in this thesis is to avoid increasing the complexity for the sentiment analysis, in the way that more parameters
(more granularities) are considered on the side of the dataset, more parameters need to be considered on the side of sentiment analysis to make proper comparisons.

However considering only one supreme question does not compromise that quality and outcome of the thesis. It will be shown in further chapters that the selected supreme question will be later a good reference point to observe how other elements affect the voting preferences of the voters during the time.

Lastly, the block with social topic questions is considered entirely. The description of each of the questions in this block is presented in the following sections by dataset.

### 4.2.1 Ecuador Presidential 2013

The 30 social topic questions have the next order in the Ecuador dataset:

| q1 | The state should monitor the economic decisions of the private sector. |
| q2 | Concessions should be provided to private foreign investment. |
| q3 | Employment opportunities should come primarily from the private sector. |
| q4 | Restricting imports is a favorable measure for the development of the country. |
| q5 | Reducing the tax burden encourages economic development. |
| q6 | The market can solve most problems more efficiently than the state. |
| q7 | Dollarization has proved negative for the development of the Ecuadorian economy. |
| q8 | The Free Trade Agreements (FTA) favour the economic development of the country. |
| q9 | A policy of greater proximity to the U.S. is beneficial to the country. |
| q10 | The withdrawal of the Manta base has brought more good than harm to Ecuador. |
| q11 | The Ecuadorian state should declare FARC to be terrorists. |
| q12 | The new Constitution has damaged the quality of Ecuadorian democracy. |
| q13 | The compulsory vote should be maintained. |
| q14 | The state should monitor the activities of the media more. |
| q15 | The optional vote of police and military personnel should be withdrawn. |
| q16 | The Prosecutor and the auditor (Contralor) should be elected by the Assembly and the Council of Citizens' Participation. |
| q17 | The national defense budget should be increased. |
| q18 | A strict public security policy should be implemented, even if it violates the human rights of offenders. |
| q19 | The increase in the number of foreigners in the country has brought about a rise in insecurity. |
| q20 | To reduce insecurity it is necessary to increase the penalties for criminals. |
| q21 | Under certain circumstances, citizens should be allowed to bear arms. |
| q22 | The government should spend more on public health, even if this means increasing taxes. |
| q23 | The government should spend more on public education, even if this means increasing taxes. |
| q24 | Access to public universities should be available without an admission examination. |
| q25 | The model of development should prioritize the protection of the environment. |
The Catholic Church is an actor that must be consulted in certain political situations.

Abortion in the early months of pregnancy should be decriminalized.

Same-sex marriages should be legalized.

Legalizing drug consumption will lead to the loss of social values.

Ideally, politicians should believe in God.

**Figure 7. Ecuador Social Issue Questions**

Each of the questions have the following scale after normalization: 
-1 = “Completely Disagree”, -0.5 = “Disagree”, 0 = “Neither agree nor disagree”, 0.5 = “Agree”, 1 = “Completely Agree”, 99 = “No opinion”

### 4.2.2 Peru Presidential 2011

In the same way as in the Ecuador case, the 30 social topic questions have the next order in the Peru’s dataset:

<table>
<thead>
<tr>
<th>q1</th>
<th>The Peruvian state, rather than the public sector, should be the owner of the most important businesses and industries of the country.</th>
</tr>
</thead>
<tbody>
<tr>
<td>q2</td>
<td>The market can resolve the problems in our society because it distributes resources in a more efficient manner than the state.</td>
</tr>
<tr>
<td>q3</td>
<td>The government should limit, by law, interest rates charged by banks.</td>
</tr>
<tr>
<td>q4</td>
<td>The government should control the prices of essential goods.</td>
</tr>
<tr>
<td>q5</td>
<td>It should be easier for companies to hire and fire employees.</td>
</tr>
<tr>
<td>q6</td>
<td>To keep unemployment rates low it would be acceptable to have a higher rate of inflation.</td>
</tr>
<tr>
<td>q7</td>
<td>To balance the budget it is better to raise taxes than to cut spending.</td>
</tr>
<tr>
<td>q8</td>
<td>The Peruvian government must honor the terms of the contracts on which foreign companies have invested in Peru.</td>
</tr>
<tr>
<td>q9</td>
<td>It is more important to encourage economic growth than to protect the environment.</td>
</tr>
<tr>
<td>q10</td>
<td>It is better to finance road construction by private investment than through taxes levied on all taxpayers.</td>
</tr>
<tr>
<td>q11</td>
<td>Do you agree with a windfall tax on mining?</td>
</tr>
<tr>
<td>q12</td>
<td>After the reduction of IGV (general sales tax) from 19% to 18%, do you think that IGV should be reduced still further?</td>
</tr>
<tr>
<td>q13</td>
<td>The government should spend more on public health services, even if this may involve raising taxes.</td>
</tr>
<tr>
<td>q14</td>
<td>The government should spend more on public education, even if this may involve raising taxes.</td>
</tr>
<tr>
<td>q15</td>
<td>Do you agree that teachers’ salaries should be increased unconditionally?</td>
</tr>
<tr>
<td>q16</td>
<td>Camisea gas should cover domestic consumption before being exported.</td>
</tr>
<tr>
<td>q17</td>
<td>The Free Trade Agreement with the United States should be renegotiated.</td>
</tr>
<tr>
<td>q18</td>
<td>Peru should make more effort towards integration with neighboring countries than in relations with the United States and Europe.</td>
</tr>
<tr>
<td>q19</td>
<td>Peru should introduce the death penalty for the rape of minors.</td>
</tr>
<tr>
<td>q20</td>
<td>The consumption of marijuana should be decriminalized in Peru.</td>
</tr>
<tr>
<td>q21</td>
<td>Homosexual couples should have the right to establish civil partnerships.</td>
</tr>
<tr>
<td>q22</td>
<td>Abortion in the early months of pregnancy should be decriminalized.</td>
</tr>
<tr>
<td>q23</td>
<td>Should Compulsory Military Service be re-introduced?</td>
</tr>
<tr>
<td>q24</td>
<td>Do you agree that the budget for the defense sector should be increased?</td>
</tr>
</tbody>
</table>
A strict public security policy should be established, even if it violates the human rights of offenders.

The state child care program (Wawa Wasi) should be expanded.

Do you agree that the salaries of senior public officials should be increased?

Should compulsory voting be maintained?

Should the Congress once again have two chambers: Deputies and senators?

Parliamentary immunity should be abolished.

Each of the questions have the following scale after normalization: -1 = “Completely Disagree”, -0.5 = “Disagree”, 0 = “Neither agree nor disagree”, 0.5 = “Agree”, 1 = “Completely Agree”, 99 = “No opinion”

4.3 Dataset Pre-Processing

Based on the dataset structure for both datasets, this thesis only contains the candidates’ vectors for the top five candidates.

In the case of Peru, there are some political party alliances, which make difficult to represent a concrete ideology or political position that is why the analysis relies rather on a candidate than a political party.

In the case of Ecuador there are not party alliances and each candidate represent the ideologies of their party, but 3 candidates in the dataset does not have enough vote intention data, which represents a size reduction dataset problem (further explained in section 4.4). As in the Peru’s case, the top 5 candidates are also considered for Ecuador’s case.

The next subsections (4.3.1 and 4.3.2) show the party alliances and the political party that each candidate represent.

4.3.1 Elections in Peru 2011 Candidates and Parties

<table>
<thead>
<tr>
<th>Political Party or Group</th>
<th>Presidential candidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alianza Gana Perú</td>
<td>Ollanta Humala</td>
</tr>
<tr>
<td>Peruvian Nationalist Candidate</td>
<td></td>
</tr>
<tr>
<td>(Partido Nacionalista Peruano)</td>
<td></td>
</tr>
<tr>
<td>Socialist Candidate</td>
<td></td>
</tr>
<tr>
<td>(Partido Socialista)</td>
<td></td>
</tr>
<tr>
<td>Peruvian Communist Candidate</td>
<td></td>
</tr>
<tr>
<td>(Partido Comunista Peruano)</td>
<td></td>
</tr>
<tr>
<td>Revolutionary Socialist Candidate</td>
<td></td>
</tr>
<tr>
<td>(Partido Socialista Revolucionario)</td>
<td></td>
</tr>
<tr>
<td>Political Movement Socialist Voice</td>
<td></td>
</tr>
<tr>
<td>(Movimiento Político Voz Socialista)</td>
<td></td>
</tr>
</tbody>
</table>
**Fuerza 2011**
- Force 2011 (Fuerza 2011)
- National Renewal (Renovación Nacional)

**Alianza Perú Posible**
- Possible Peru (Perú Posible)
- We Are Peru (Somos Perú)
- Popular Action (Acción Popular)

**Alianza por el Gran Cambio**
- Alliance for Progress (Alianza para el Progreso)
- Humanist Candidate (Partido Humanista)
- Christian People’s Candidate (Partido Popular Cristiano)
- National Restoration (Restauración Nacional)

**Alianza Solidaridad Nacional**
- Change 90 (Cambio 90)
- National Solidarity (Solidaridad Nacional)
- Always Together (Siempre Unidos)
- Union for Peru (Unión por el Perú)

*Figure 9. Elections in Peru 2011 Candidates and Parties*

### 4.3.2 Elections in Ecuador 2013 Candidates and Parties

<table>
<thead>
<tr>
<th>Political Party</th>
<th>Candidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movimiento Alianza PAIS</td>
<td>Rafael Correa</td>
</tr>
<tr>
<td>Creating Opportunities</td>
<td>Guillermo Lasso</td>
</tr>
<tr>
<td>January 21 Patriotic Society Party</td>
<td>Lucio Gutiérrez</td>
</tr>
<tr>
<td>SUMA</td>
<td>Mauricio Rodas</td>
</tr>
<tr>
<td>Institutional Renewal Party of National Action</td>
<td>Álvaro Noboa</td>
</tr>
<tr>
<td>Plurinational Unity of the Lefts</td>
<td>Alberto Acosta</td>
</tr>
<tr>
<td>Ruptura 25</td>
<td>Norman Wray</td>
</tr>
<tr>
<td>Ecuadorian Roldosist Party</td>
<td>Nelson Zavala</td>
</tr>
</tbody>
</table>

*Figure 10. Elections in Ecuador 2013 Candidates and Parties*
4.4 Dataset Processing

The golden rule for datasets is always analyze the data in terms of completeness, which means to consider the voters that have answered most of the questions in the survey. In order to clean the data set and remain with a proper set of users, it is needed to reduce the dataset size and also provide fairness in terms of amount of users because of the nature of fuzzy algorithm that will be used later during the evaluation chapter. The data set reduction for both cases was performed in 4 steps. Figure 11 illustrates Steps 1 to Step 4 of the cleaning process.

- **Step 1:** The analysis only consider voters who fully answered all the 30 issue statements, i.e. any user that had at least one “99” value was exclude from the analysis. The reason is that the research model in this thesis is designed to only use fully answered questionnaires.

- **Step 2:** The aim of this step was to downsize the datasets with users who also gave a clear statement about their vote intention. Users who answered “Other”, “None” or “Did not supply information” were excluded from the dataset.

- **Step 3:** In this step, the dataset reduction is based on the voter’s amount per candidate. The equal distribution of number of users per candidate is crucial because if fairness is not given, the Fuzzy C-mean Algorithm would give
preference to the candidate who has more voters. Then both datasets were reduced in order to have an equal amount of voters among candidates.
  
  - Peru with 500 voters per candidate
  - Ecuador with 424 voters per candidate

- **Step 4:** The last step is related to the dimensionality reduction of the dataset. Originally, the candidates’ vector for the candidates contained 30 issue questions. To get a clearer image of the political landscape, only fully answered issue questions by the candidates must be taken into account for later analysis. Then based on the 30 social topic questions of the candidate answers, an elimination of incomplete answers with no opinion were removed in the dataset as well.

The next Figure 12 shows the effects of the dataset processing in each step of the cleaning process.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Original Size</th>
<th>Phase 1 Size</th>
<th>Phase 2 Size</th>
<th>Phase 3 Size</th>
<th>Phase 4 Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peru</td>
<td>30 x 40627</td>
<td>30 x 32717</td>
<td>30 x 26149</td>
<td>30 x 2500</td>
<td>28 x 2500</td>
</tr>
<tr>
<td>Ecuador</td>
<td>30 x 11531</td>
<td>30 x 8586</td>
<td>30 x 6030</td>
<td>30 x 2120</td>
<td>30 x 2120</td>
</tr>
</tbody>
</table>

*Figure 12. Dataset Processing for Peru and Ecuador*
5. Sentiment Analysis

Sentiment analysis, or opinion mining, is an automated knowledge discovery technique that aims to find hidden patterns in a large amount of textual information, including social media (Mostafa, 2013). The goal of sentiment analysis is to create a knowledge base in a structured, explicit manner containing reviews (positive, negative, and neutral) that express sentiments, evaluations, and perceptions about any subject (Fortuny, Smedt, Martens, & Daelemans, 2012; Sobkowicz, Kaschesky, & Bouchard, 2012).

In this context, this study aims to use some sentiment analysis as a tool to allow our candidate model to use information circulating on social media for the evaluation of social topics. In other words, this study describes how sentiment analysis can be used as an instrument for measuring public opinion of the candidates about social topics and identifying citizens’ main dissatisfactions with the political parties, so that a dynamic candidate profile can be created. The next sub-section of this chapter presents the theoretical dynamic candidate model.

5.1 Sentiment Analysis Elements

The social media platform that is considered in this study to perform sentiment analysis is Twitter, which is an online social networking service that enables users to send and read short 140-character messages. The advantage of these social media is that all the posts are public and it is possible to recall them through an API. The Figure 13 shows the categories to be considered as part of the sentiment analysis architecture and their connection among them.

The sentiment analysis considers two main categories involved in the construction of the dynamic part of the candidate’s profile:

- **Candidate Posts:** The messages that the candidate retweet or write personally in his/her account.

- **Candidate Tags:** It represents the comments from other twitter users (voters) where the candidate may be related.
The Dynamic candidate’s vector is built by the sum of two main categories (Candidate Tags and Candidate Posts), which are mapped between the tweets and social categories. The equation 1.0 represents the calculation of the Dynamic Candidate’s Vector, where $a$ is a fraction between 1 and 0. For instance if the value of $a$ is 0.3 then the candidate posts have 30% of importance and the tags from the users 70%. This creates a more realistic scenario to create a vector that represents the voters’ candidate perception.

$$\sum_i A_i IV_i a + B_i IV_i (1 - a)$$ (1.0)

The tweets streaming capture is performed automatically by a python program, part of this thesis research, created to read and count the words in the tweets of each candidate. Later with a Spanish sentiment analysis dictionary and human analysis, it is possible to classify the words by topic and assign values in the vectors.
As it was mentioned before, the value of a ensures that the two main categories do not have the same relevance, for instance the candidate posts may be fictitious or manipulated in order to keep a good self-image, so it can be biased by the candidate. However, the posts or tweets where the candidate has been tagged are more authentic and representative but there might be still the problem that several external users can work cooperatively to affect the public image of a candidate by posting compromising messages or an internet bot.

At the end, the sentiment analysis model considers that the users who tag a candidate are authentic by observing the number of followers that the user has. Most of the Internet bots do not have followers because the accounts are fake and automatically created. However it is out of the scope of this thesis to find an ideal method to avoid this potential bias but it tries to reduce it.

The model by default considers an importance vector also, which gives more importance to the tweets that come from the twitter users rather than the ones from the candidate. On the other hand, this study presents in the analysis section the level of candidate popularity in social networks by counting the number of followers.

Finally this dynamic vector can be calculated more often and it can help to the users in the way that it represents in a more spontaneous way the candidate perception, and in terms of campaign strategy, it reflects the strengths and weaknesses of the candidate that may be improved during a political campaign.

5.2. Mapping Tweets to a Vector

Based on the figure 13, in order to obtain the initial mapped vectors it was needed to create and adapt several programs in python that could extract in a stream the tweets of the candidates with certain filters till create a vector that represents the twitts in the social topic.

The Figure 14 shows the entire process and it started with a python program that capture tweets and save them in a CVS file, later another Python program classifies the tweets by users and eliminate the emoticons, URLs and other non-relevant symbols and perform a word counter per candidate.

Once the most important words were obtained per tweet and candidate, it was necessary to assign a weight by a sentiment spanish dictionary. However the outcome
was not perfect, thus human intervention was needed to analyse the outcome of the
dictionary evaluation and map the word in a social category.

For instance the word “revolution” was a top word from one of the candidates and the
outcome of the dictionary was a positive rate, nevertheless this is false because the
word coming from the candidate was in a negative sense. Therefore human
intervention was needed to assign the right negative value and classify the word into a
social category.

5.3 Sentiment Spanish Dictionary

Because of the lack of a spanish sentiment dictionary, the sentiment dictionary used
in this thesis was an adaptation of a current english dictionary used already for this
purposes. The AFINN dictionary is a list of English words rated for values with an
integer between minus five (negative) and plus five (positive). The words have been
manually labeled by Finn Årup Nielsen in 2009-2011 at the Technical University of
Denmark, the dictionary contains 2477 words and phrases.

The construction of the dictionary took three steps, the first part was to use a translator,
in this case google translate to convert the words from english to spanish, the phrases
were discarded because of the inaccurate translation by google and also because the
local spanish expressions variations will not match these phrases. The second part
involved to change the rates of the weights into a scale that is compatible with our
profile vectors. The figure 15 shows the equivalence in values.
The third part involved to analyse again the dictionary words in order to adapt it to the Spanish dialects of Peru and Ecuador. Although in Latin America, Spanish is the most spoken language, the use of words is different and every culture has different words in politics that most of the time are not the same. Therefore the final Spanish dictionary contained around 100 Spanish useful words among the total words translated for Peru and Ecuador.

At the end the dictionary creation, helped to assign weights to the Spanish words from the tweets but it was not perfect and in most of the cases it performed poorly, therefore as it was mentioned in the previous section, human intervention was needed to understand the context of the words and in some cases change the weight of the words manually.
6. Voting Advice Application Algorithms

Knowing all the parts that compose our candidate profile model, it is time to explain some algorithms in which our candidate model can be analyzed.

The methods or algorithms used for calculating the policy congruence in different studies and papers previously consulted showed that there is not a standard algorithm to apply because all of them differ substantially between VAAs. While some use Euclidian distance to find the closest match, others use the City Block model. These two distinct mathematical formulas for calculating distances between objects of interest (see Louwerse & Rosema, 2011) affect the results of the matching procedure and therefore the voting recommendation of the VAA.

VAA designs are currently under intense scientific scrutiny, especially in terms of the matching algorithms they employ (Gemenis, 2012). Last but not least, the presentation of results varies between VAAs, with various different procedures for how to visualize several issue positions in a reduced form. VAAs are increasingly used in election campaigns worldwide, thus their design and methodology deserve closer attention, especially because the outputs they produce seem to affect those who use them.

In this chapter and for the purposes of our study, two algorithms are considered: Euclidian distance and a Fuzzy C means Algorithm. The theory behind these algorithms is explained in the next sections. Once their concept and application is clear, then the theoretical background is covered in order to proceed with the case analysis in subsequent chapters.

6.1 Euclidean distance

The simplest approach to measure similarity is the Euclidean distance, where $d(x,y)$ is the degree of the distance:

$$d(x,y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \quad (1.1)$$

Where $n$ is the number of dimensions (attributes) and $x_k$ and $y_k$ are the $k^{th}$ attributes (components) of data objects $x$ and $y$, respectively [6].
6.2 Fuzzy C-means Algorithm

The Fuzzy C-means clustering algorithm is based on the minimization of an objective function called C-means functional. It is defined by Dunn as:

\[ J(X; U, V) = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^m \| x_i - v_k \|_A^2 \]  \hspace{1cm} (1.2)

Where:

\[ V = [v_1, v_2, \ldots, v_c], \quad v_i \in \mathbb{R}^n \]  \hspace{1cm} (1.3)

is a vector of cluster prototypes (centers), which have to be determined, and

\[ D_{ik}^2 = \| x_i - v_k \|_A^2 = (x_i - v_k)^T A (x_i - v_k) \]  \hspace{1cm} (1.4)

is a squared inner-product distance norm.

Statistically, (1.2) can be seen as a measure of the total variance of \( x_k \) from \( v_i \). The minimization of the c-means functional (1.2) represents a nonlinear optimization problem that can be solved by using a variety of available methods, ranging from grouped coordinate minimization, over simulated annealing to genetic algorithms. The most popular method, however, is a simple Picard iteration through the first-order conditions for stationary points of (1.2), known as the fuzzy c-means (FCM) algorithm.

The stationary points of the objective function (1.2) can be found by adjoining the constraint (1.5) to \( J \) by means of Lagrange multipliers (1.6):

\[ \sum_{k=1}^{c} \mu_{ik} = 1, \quad 1 \leq i \leq N \]  \hspace{1cm} (1.5)

\[ J(X; U, V, \lambda) = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^m D_{ik}^2 + \sum_{k=1}^{N} \lambda_k \left( \sum_{i=1}^{c} \mu_{ik} - 1 \right) \]  \hspace{1cm} (1.6)

and by setting the gradients of (\( \bar{J} \)) with respect to \( U, V \) and \( \lambda \) to zero. If \( D_{ik} > 0, \forall i, k \) and \( m > 1 \), then \( (U, V) \in \mathbb{M}_c \times \mathbb{R}^{nc} \) may minimize (1.2) only if

\[ \mu_{ik} = \frac{1}{\sum_{j=1}^{c} \left( \frac{D_{ik}}{D_{jk}} \right)^{2/(m-1)}}, \quad 1 \leq i \leq c, 1 \leq k \leq N, \]  \hspace{1cm} (1.7)
And

\[ v_i = \frac{\sum_{k=1}^{N} \mu_{ik}^m x_k}{\sum_{k=1}^{N} \mu_{ik}^m}, \quad 1 \leq i \leq c. \]  \hspace{1cm} (1.8)

Note that equation (1.8) gives \( v_i \) as the weighted mean of the data items that belong to a cluster, where the weights are the membership degrees. That is why the algorithm is called "c-means". One can see that the FCM algorithm is a simple iteration through (1.7) and (1.8).

The FCM algorithm computes with the standard Euclidean distance norm, which induces hyper spherical clusters. Hence it can only detect clusters with the same shape and orientation, because the common choice of norm inducing matrix is: \( A = I \) or it can be chosen as an \( n \times n \) diagonal matrix that accounts for different variances in the directions in the directions of the coordinate axes of \( X \):

\[
A_{D} = \begin{bmatrix}
(1/\sigma_1)^2 & 0 & \cdots & 0 \\
0 & (1/\sigma_2)^2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & (1/\sigma_n)^2
\end{bmatrix} \quad (1.9)
\]

or \( A \) can be defined as the inverse of the \( n \times n \) covariance matrix: \( A = F^{-1} \), with

\[
F = \frac{1}{N} \sum_{k=1}^{N} (x_k - \bar{x})(x_k - \bar{x})^T. \quad (1.10)
\]

Here \( \bar{x} \) denotes the sample mean of the data. In this case, \( A \) induces the Mahalanobis norm on \( \mathbb{R}^n \).
ALGORITHM STEPS:
Given the data set $X$, choose the number of clusters $1 < c < N$, the weighting exponent $m > 1$, the termination tolerance $\varepsilon > 0$ and the norm-inducing matrix $A$. Initialize the partition matrix randomly, such that $U^{(0)} \in M_{fc}$.

Repeat for $l = 1, 2, \ldots$

Step 1 Compute the cluster prototypes (means):

$$v_i^{(l)} = \frac{\sum_{k=1}^{N} (\mu_{i,k}^{(l-1)})^{m} x_k}{\sum_{k=1}^{N} (\mu_{i,k}^{(l-1)})^{m}}, \quad 1 \leq i \leq c.$$  \hfill (1.11)

Step 2 Compute the Distances:

$$D_{ikA}^{2} = (x_k - v_i)^T A (x_k - v_i), \quad 1 \leq i \leq c, \quad 1 \leq k \leq N.$$  \hfill (1.12)

Step 3 Update the partition matrix:

$$\mu_{i,k}^{(l)} = \frac{1}{\sum_{j=1}^{c} (D_{ikA}/D_{jA})^{2/(m-1)}},$$  \hfill (1.13)

until $\|U^{(l)} - U^{(l-1)}\| < \varepsilon$.

6.3 Principal Component Analysis (PCA)

Las but not least, one statistical procedure that helps us to perform later the algorithms evaluation is the principal component analysis (PCA), which involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The main objectives of PCA are:

1. Identify new meaningful underlying variables.
2. Discover or to reduce the dimensionality of the data set.
The mathematical background lies in "Eigen analysis": The eigenvector associated with the largest eigenvalue has the same direction as the first principal component. The eigenvector associated with the second largest eigenvalue determines the direction of the second principal component.

In this thesis, we used the second objective, in that case the covariance matrix of the data set (also called the "data dispersion matrix") is defined as follows:

\[
F = \frac{1}{N} (x_k - \mu) (x_k - \mu)^T 
\]

(1.15)

Where \( \mu = \overline{x_k} \), the mean of the data (\( N \) equals the number of objects in the data set).

Principal Component Analysis (PCA) is based on the projection of correlated high-dimensional data onto a hyperplane. This mapping uses only the first few \( q \) nonzero eigenvalues and the corresponding eigenvectors of the \( F = U \Lambda U^T \), covariance matrix, decomposed to the \( \Lambda \) matrix that includes the eigenvalues \( \lambda_{i,j} \) of \( F \) in its diagonal in decreasing order, and to the \( U \) matrix that includes the eigenvectors corresponding to the eigenvalues in its columns. The vector \( y_{i,k} = W^{-1}_i(x_k) = W^T_i(x_k) \) is a \( q \)-dimensional reduced representation of the observed vector \( x_k \), where the \( W_i \) weight matrix contains the \( q \) principal orthonormal axes in its column \( W_k = U_{i,q} \Lambda_{i,q}^{\frac{1}{2}} \).

It sounds complicated the first time but it will become clearer with some interpretation examples that are introduced in the next chapter.
7. Analysis and Evaluation

In this chapter, the analysis will focus mostly on the evaluation the candidate profile and its application with two different VAA’s algorithms. Given that the analysis is extensive in this chapter and it is easy to lose the track, the Figure 16 and Figure 17 present two different scenarios with all the stages that are going to be covered in this chapter.

Figure 16. Scenario A without Sentiment Analysis

Figure 17. Scenario B with Sentiment Analysis
The whole analysis in this chapter is always based on a comparison between these two main scenarios:

- Scenario A: No sentiment analysis is used to build the candidate’s vector and the study specially test the Euclidian and the Fuzzy Clustering algorithms.
- Scenario B: Sentiment analysis plays a main role in the candidate’s profile and the Fuzzy Clustering algorithm is tested.

The first subsection (7.1) in the document, shown in Figure 17, starts with the construction of the dynamic vector by sentiment analysis. Later the candidate vector (7.2) is created and integrated in the dataset (7.4). Posterior, the data is computed with the fuzzy algorithm (7.5), in order to obtain the degrees of membership (7.6) per voter. In parallel, a small analysis by Euclidian distance is performed and compared for both scenarios between the candidates’ vectors and their integration with the dataset (7.3).

7.1 Sentiment Analysis

This part is focused to describe the methodology and explain relevant statistics for the dynamic part of the candidate’s profile.

7.1.1 Statistics of Twitter Accounts

There are several statistics or altmetrics that can be calculated and obtained by eavesdropping the data of Twitter users. Particularly the presence of the candidates in the social network Twitter. The next altmetrics helped the study to determine which candidates should be considered for the study and for the generation of a candidate profile. It will be explained later but some candidates do not have enough information or present an absence of information to perform sentiment analysis and eventually later the VAA algorithms, as it will be observed in the case of Ecuador.

7.1.1.1 Ecuador

During the study, it was possible to obtain the number of followers and the amount of Tweets that each candidate has since the creation of their accounts. This will help to grasp an idea about their internet activity.

The number of followers is presented in the Figure 18. The top candidate is Rafael Correa with 2732486 followers, this represents a huge gap with the other candidates, especially with Pastor Nelson Zavala who barely has followers (842).
Another relevant altmetric is the number of tweets that each candidate publish. Figure 19 shows who is more active in the twitter account and it is not necessary the one who has more followers.

These altmetrics are just the overview of several options, it is possible also to get more granularity such as classification of followers by sex, location, etc. Nevertheless in the case of Ecuador only the top 5 candidates were considered because of the lack of mentions or activity in some candidates in Twitter.
7.1.1.2 Peru

The number of followers for the case of Peru is represented by the Figure 20, where it is possible to appreciate that the most popular candidate is Ollanta Humala with 1427735 followers and the least popular Luis Castañeda with 19044 followers. It is important to mention that for the case of Peru, also the top 5 candidates are considered for the analysis but there are more that were not considered because of lack of information previously explained in previous sections.

![Figure 20. Followers of Peru Candidates](image)

The candidate activity by Tweets is presented in the Figure 21 and it shows that Pedro Pablo Kuczynski is the most active in the social network.

![Figure 21. Number of Tweet Status for Peru Candidates](image)
All these statistics can be understood in many ways and more granularity is needed in order to provide a proper interpretation but in our case it shows that they can be a good factor that influence the VAA’s design methodology.

7.1.1.3 User Tweets

Extracting the tweets from an account has a limit in twitter. A total of 3200 tweets can be captured with recently libraries and then perform lately the word count. As an example the Figure 22 and Figure 23 show the top words among the candidates.

<table>
<thead>
<tr>
<th>Toledo</th>
<th>Fujimori</th>
<th>Castañeda</th>
</tr>
</thead>
<tbody>
<tr>
<td>Para Volver a Crecer</td>
<td>Humala</td>
<td>Lima</td>
</tr>
<tr>
<td>Gobierno</td>
<td>Gobierno</td>
<td>simplementetrabajar do</td>
</tr>
<tr>
<td>democracia</td>
<td>JuntosHaciaElFuturo</td>
<td>alcalde</td>
</tr>
<tr>
<td>seguridad</td>
<td>Peruanos</td>
<td>Obras</td>
</tr>
<tr>
<td>Crecimiento</td>
<td>democracia</td>
<td>Trabajar</td>
</tr>
<tr>
<td>Propuestas</td>
<td>Politica</td>
<td>Campana</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Kuczynski</th>
<th>Humala</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unidos la hacemos</td>
<td>Ganaperu</td>
</tr>
<tr>
<td>ppkamigo</td>
<td>gobierno</td>
</tr>
<tr>
<td>propuestas</td>
<td>compatriotas</td>
</tr>
<tr>
<td>gobierno</td>
<td>compromiso</td>
</tr>
<tr>
<td>agua</td>
<td>campana</td>
</tr>
<tr>
<td>economia</td>
<td>peruanos</td>
</tr>
</tbody>
</table>

Figure 22. Peru’s Candidate Tweets - Top Words

<table>
<thead>
<tr>
<th>Mauricio Rodas</th>
<th>Guillermo Lasso</th>
<th>Norman Wray</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quiteños</td>
<td>Entrevista</td>
<td>Propuesta</td>
</tr>
<tr>
<td>construcción</td>
<td>Por el Empleo</td>
<td>Rupturaec</td>
</tr>
<tr>
<td>ayuda</td>
<td>Democracia</td>
<td>entrevista</td>
</tr>
<tr>
<td>prevención</td>
<td>crisis</td>
<td>mujeres</td>
</tr>
<tr>
<td>solidaridad</td>
<td>inversion</td>
<td>alegria</td>
</tr>
<tr>
<td></td>
<td>Words</td>
<td>Amount</td>
</tr>
<tr>
<td>--------</td>
<td>-----------</td>
<td>--------</td>
</tr>
<tr>
<td>Correa</td>
<td></td>
<td>261</td>
</tr>
<tr>
<td>Gobierno</td>
<td></td>
<td>139</td>
</tr>
<tr>
<td>Ecuador</td>
<td></td>
<td>131</td>
</tr>
<tr>
<td>correismo</td>
<td></td>
<td>66</td>
</tr>
<tr>
<td>derechos</td>
<td></td>
<td>63</td>
</tr>
</tbody>
</table>

*Figure 23. Ecuador’s Candidate Tweets - Top Words*

It is interesting that by just observing the words, it is possible to have a feeling or a sentiment for the personality of the candidate. Although by observing the word is very subjective and more context is needed, it is interesting to notice a certain style in their way to communicate their ideas with the voters.
7.1.1.4 Tagged

The python programs also allow us to extract streams and be able to capture other tweets that were published by other users who can tag a candidate and communicate his/her ideas to him.

The capture of the streams was performed at random times during a week, the raw data that was used in this master thesis is provided additionally and the capture time can be consulted directly there.

The eavesdropping was performed during a week at random times for one hour or one day or per hour. Since capturing tweet streams do not have buffer, sometimes it was necessary to repeat the test because no relevant tweets were captured. Therefore the raw data for this study was the result of several failure attempts and only the successful data was collected among many different samples. The idea behind of capturing in a random way, has the aim to catch different commentaries about political events. For instance the Ecuador earthquake disaster was captured partially and it deviated the attention of the users and it was not possible to map these tweets to the social topics analyzed in this thesis. As a consequence it was needed to wait some days in order to capture more relevant tweets.

In addition finding useful words that are related to the social topics is very hard because normally when users tag a candidate, they normally send pictures, emoticons or local and short Spanish expressions that cannot be in the majority of the cases classified or mapped in the political issues. Some examples of such words are “Corrupto”, “Mentiroso”, “Ladrón”, etc. However this mapping was also done as in the previous case by human intervention and understanding the context of the entire tweets rather than focusing in the amount of words. The Figure 24 shows the twitter references of the candidates during the monitoring time.

<table>
<thead>
<tr>
<th>Candidate</th>
<th>Twitter References</th>
<th>Candidate</th>
<th>Twitter References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mauricio Rodas</td>
<td>77</td>
<td>Toledo</td>
<td>78</td>
</tr>
<tr>
<td>Guillermo Lasso</td>
<td>94</td>
<td>Fujimori</td>
<td>10000</td>
</tr>
<tr>
<td>Norman Wray</td>
<td>5</td>
<td>Castañeda</td>
<td>79</td>
</tr>
<tr>
<td>Alberto Acosta</td>
<td>6</td>
<td>Kuczynski</td>
<td>7960</td>
</tr>
<tr>
<td>Rafael Correa</td>
<td>3890</td>
<td>Humala</td>
<td>836</td>
</tr>
</tbody>
</table>

Figure 24. Twitter References
Based on all this information and the context of the analysis, it was possible to create the dynamic part of the candidate’s profile.

7.1.2 Constructing the Dynamic Vector

The aim of this section is to present the results of the dynamic part of the candidate’s profile. The Figure 13 showed previously the process to build this part of the profile. The results that are shown here represent only the top 5 candidates of each country because the other candidates do not have enough information to be considered or they remain inactive in social networks after the elections period in 2013 and 2011 for Ecuador and Peru respectively.

7.1.2.1 Ecuador

The dynamic vector for Ecuador is presented in the next Figure 25

<table>
<thead>
<tr>
<th></th>
<th>Mauricio Rodas</th>
<th>Guillermo Lasso</th>
<th>Norman Wray</th>
<th>Alberto Acosta</th>
<th>Rafael Correa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>-0.5000</td>
<td>-1.0000</td>
<td>0.5000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>Q2</td>
<td>1.0000</td>
<td>1.0000</td>
<td>-0.5000</td>
<td>-0.5000</td>
<td>-0.1500</td>
</tr>
<tr>
<td>Q3</td>
<td>0.5000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>-0.5000</td>
<td>-0.5000</td>
</tr>
<tr>
<td>Q4</td>
<td>-0.5000</td>
<td>-1.0000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>1.0000</td>
</tr>
<tr>
<td>Q5</td>
<td>0.5000</td>
<td>1.0000</td>
<td>-0.5000</td>
<td>-1.0000</td>
<td>-0.5000</td>
</tr>
<tr>
<td>Q6</td>
<td>0.5000</td>
<td>0.0000</td>
<td>-0.5000</td>
<td>-1.0000</td>
<td>-1.0000</td>
</tr>
<tr>
<td>Q7</td>
<td>-0.5000</td>
<td>-1.0000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>1.0000</td>
</tr>
<tr>
<td>Q8</td>
<td>0.5000</td>
<td>1.0000</td>
<td>-0.5000</td>
<td>-1.0000</td>
<td>-1.0000</td>
</tr>
<tr>
<td>Q9</td>
<td>0.5000</td>
<td>0.5000</td>
<td>-0.5000</td>
<td>-1.0000</td>
<td>-1.0000</td>
</tr>
<tr>
<td>Q10</td>
<td>-0.5000</td>
<td>-0.5000</td>
<td>0.5000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>Q11</td>
<td>0.5000</td>
<td>0.5000</td>
<td>-1.0000</td>
<td>-1.0000</td>
<td>-1.0000</td>
</tr>
<tr>
<td>Q12</td>
<td>1.0000</td>
<td>0.5000</td>
<td>-1.0000</td>
<td>-1.0000</td>
<td>-0.5000</td>
</tr>
<tr>
<td>Q13</td>
<td>-0.5000</td>
<td>-0.5000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.5000</td>
</tr>
<tr>
<td>Q14</td>
<td>-0.1500</td>
<td>-1.0000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>1.0000</td>
</tr>
<tr>
<td>Q15</td>
<td>0.0000</td>
<td>0.0000</td>
<td>-0.5000</td>
<td>-0.5000</td>
<td>-1.0000</td>
</tr>
<tr>
<td>Q16</td>
<td>0.5000</td>
<td>0.5000</td>
<td>-0.5000</td>
<td>-0.5000</td>
<td>-1.0000</td>
</tr>
<tr>
<td>Q17</td>
<td>0.0000</td>
<td>0.5000</td>
<td>-0.5000</td>
<td>-0.5000</td>
<td>1.0000</td>
</tr>
<tr>
<td>Q18</td>
<td>-1.0000</td>
<td>0.0000</td>
<td>-1.0000</td>
<td>-1.0000</td>
<td>-0.5000</td>
</tr>
<tr>
<td>Q19</td>
<td>-0.5000</td>
<td>0.5000</td>
<td>-1.0000</td>
<td>-1.0000</td>
<td>-0.5000</td>
</tr>
<tr>
<td>Q20</td>
<td>-0.8500</td>
<td>0.0000</td>
<td>-1.0000</td>
<td>-1.0000</td>
<td>0.5000</td>
</tr>
</tbody>
</table>
### 7.1.2.1 Peru

As it was discussed in section 4.4, Peru’s questions were reduced to 28 values instead of 30 because of a lack of the answer of some candidates. The final dynamic vector for Peru is presented in Figure 26.

![Figure 25. Ecuador Dynamic Candidates’ Vectors](image_url)

**Figure 25. Ecuador Dynamic Candidates’ Vectors**

<table>
<thead>
<tr>
<th>Q21</th>
<th>0.5000</th>
<th>0.5000</th>
<th>-0.5000</th>
<th>-0.5000</th>
<th>-1.0000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q22</td>
<td>0.5000</td>
<td>0.8500</td>
<td>0.5000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>Q23</td>
<td>0.5000</td>
<td>0.8500</td>
<td>0.5000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>Q24</td>
<td>-0.5000</td>
<td>-1.0000</td>
<td>-0.5000</td>
<td>1.0000</td>
<td>-1.0000</td>
</tr>
<tr>
<td>Q25</td>
<td>0.6500</td>
<td>-0.5000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>-0.5000</td>
</tr>
<tr>
<td>Q26</td>
<td>0.5000</td>
<td>0.5000</td>
<td>-1.0000</td>
<td>-1.0000</td>
<td>-1.0000</td>
</tr>
<tr>
<td>Q27</td>
<td>-0.5000</td>
<td>-1.0000</td>
<td>1.0000</td>
<td>0.5000</td>
<td>-1.0000</td>
</tr>
<tr>
<td>Q28</td>
<td>-0.5000</td>
<td>-1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>-0.5000</td>
</tr>
<tr>
<td>Q29</td>
<td>0.5000</td>
<td>1.0000</td>
<td>-1.0000</td>
<td>-0.5000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Q30</td>
<td>0.5000</td>
<td>1.0000</td>
<td>-1.0000</td>
<td>-1.0000</td>
<td>0.5000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Toledo</th>
<th>Fujimori</th>
<th>Castaneda</th>
<th>Kuczynski</th>
<th>Humala</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Q2</td>
<td>-1</td>
<td>0.5</td>
<td>-0.5</td>
<td>0</td>
</tr>
<tr>
<td>Q3</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>Q4</td>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>Q5</td>
<td>0.5</td>
<td>0.5</td>
<td>-0.5</td>
<td>-1</td>
</tr>
<tr>
<td>Q6</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>Q7</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>Q8</td>
<td>-0.5</td>
<td>-0.5</td>
<td>-0.5</td>
<td>0</td>
</tr>
<tr>
<td>Q9</td>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>Q10</td>
<td>-1</td>
<td>0</td>
<td>-0.5</td>
<td>-1</td>
</tr>
<tr>
<td>Q11</td>
<td>-0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>Q12</td>
<td>0.5</td>
<td>0.5</td>
<td>-1</td>
<td>-0.5</td>
</tr>
<tr>
<td>Q13</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>Q14</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Q15</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>Q16</td>
<td>-1</td>
<td>-1</td>
<td>-0.5</td>
<td>0</td>
</tr>
<tr>
<td>Q17</td>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>
7.2 Candidates Vectors

Recalling the Figure 3, it is time to create the Candidates Vectors that consider the dynamic and static categories. For the purpose of this study all the social topics are considered for the analysis, so the last importance vector mentioned in the Figure 3 was considered as a vector with only 1’s.

7.2.1 Ecuador

The Ecuador candidates’ vectors provided by the experts are shown in the Figure 27.
### 7.2.2 Peru

The Peru candidates' vectors are shown in the Figure 28

<table>
<thead>
<tr>
<th>Toledo</th>
<th>Fujimori</th>
<th>Castaneda</th>
<th>Kuczynski</th>
<th>Humala</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Q2</td>
<td>-0.25</td>
<td>0.50</td>
<td>-0.50</td>
<td>0.00</td>
</tr>
<tr>
<td>Q3</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>1.00</td>
</tr>
<tr>
<td>Q4</td>
<td>0.75</td>
<td>0.50</td>
<td>0.50</td>
<td>1.00</td>
</tr>
<tr>
<td>Q5</td>
<td>0.50</td>
<td>0.50</td>
<td>-0.50</td>
<td>-1.00</td>
</tr>
<tr>
<td>Q6</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>1.00</td>
</tr>
<tr>
<td>Q7</td>
<td>0.50</td>
<td>0.00</td>
<td>0.50</td>
<td>1.00</td>
</tr>
<tr>
<td>Q8</td>
<td>-0.50</td>
<td>-0.50</td>
<td>-0.50</td>
<td>-0.50</td>
</tr>
<tr>
<td>Q9</td>
<td>1.00</td>
<td>0.50</td>
<td>0.50</td>
<td>1.00</td>
</tr>
<tr>
<td>Q10</td>
<td>-1.00</td>
<td>0.00</td>
<td>-0.50</td>
<td>-1.00</td>
</tr>
<tr>
<td>Q11</td>
<td>-0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>1.00</td>
</tr>
<tr>
<td>Q12</td>
<td>0.50</td>
<td>0.50</td>
<td>-1.00</td>
<td>-0.75</td>
</tr>
</tbody>
</table>
At this point of the master thesis, the first important part of the analysis milestone is achieved. The candidate profile will allow us to know if the sentiment analysis plays an important role or creates an impact in the results of the Fuzzy Clustering and Euclidian Distance Algorithms.
7.3 Euclidian Algorithm

The most often used algorithm and encountered space in several VAA’s recommender systems is Euclidian. The distance computation is widely used in several studies to provide a quick comparison between users or provide recommendations. The smallest the value of the distance, the greater the similarity between two elements.

In our case every question represents a dimension and all the dimensions are represented by a vector, in other words the length of the vector denotes the number of dimensions. In this section the thesis bases the analysis on the people who answered the supreme questions in the dataset about the vote intention and then it measures the euclidian distance between the voter and classified it into a candidate’s set. The Figure 29 shows the process of the algorithm.

![Figure 29. Euclidian Distance Process](image-url)
### 7.3.1 Ecuador

With a sample of 5921 voters among the top 5, the vote intention was filtered and the results for the Ecuador dataset are observed in the Figure 30.

**Figure 30. Vote Intention (Sup Question 3)**

As it is noticed, Rafael Correa is the most popular based on the supreme question with a 43%, nevertheless based on the voters answers and comparing the distance between candidates and voters the situation is totally different and Nornam Wray would have been the most popular choice with 40%. (Figure 31)

**Figure 31. Euclidian Distance – Ecuador**
Finally, considering the sentiment analysis profile vector, it is interesting to observe in figure 32 that the popularity of Rafael Correa is not as good as in previous scenarios with barely 2%. On the other hand, Mauricio Rodas would have won the elections, if they would have made it at the time of this sentiment analysis. Nevertheless, since he does not have too much activity on Twitter (See Figure 24), his original static profile vector was not affected at all, so the Euclidian distance basically took the static vector provided by the experts, which eventually could have been biased. As shown in the seminar work (Jose A. Mancera, Bossard)

![Figure 32. Euclidian Distance (Sentiment Analysis) - Ecuador](image-url)
7.3.2 Peru

With a sample of 26149, the vote intention shows in the Peru dataset that the candidate Pedro Pablo Kuczynski was the most popular. (Figure 33)

Figure 33. Vote Intention (Sup Question 5)

In contrast, when the distance between voters and candidates is performed, then Ollanta Humala was the most popular and it was eventually the final winner of that presidential elections period. (Figure 34)

Figure 34. Euclidian Distance – Peru
For the sentiment analysis case, it is thrilling to observe in Figure 35 that Keiko Fujimori and Pablo Kuczynski become very popular. It is important to mention that Peru has presidential elections in 2016, and if the elections were taken place the day of this analysis, probably both candidates would be almost equal in votes. Which was a fact that actually happened.

Finally, it seems that Ollanta Humala increased his popularity based on the original dataset, but in real life, his popularity sinks every year. This may be one of the effects that the original base dataset is static. At the end Euclidian distance can not always provide all the answers and more complex algorithms need to be taken into account.
7.4 Dataset and Candidate Vector Integration

As seen previously, euclidian distance is limited to provide some answers or details about the voters and candidates. In this part the Scenario A and Scenario B, Figure 16 and Figure 17 respectively are going to be compared in order to see the differences. In this section we present how the data looks like before the fuzzy algorithm computes the final center clusters and it is possible to appreciate already some differences.

7.4.1 Ecuador

The Figure 36 shows the position of the candidates based on the 30 social questions from Figure 7, this after we have applied the PCA algorithm to create dimensionality reduction in two dimensions. The voters are also plotted so it is possible to see them.

In contrast Figure 37 shows the initial position of each candidates but now with the proposed candidate profile. Aparently the candidates position seems the same in the superior part of the figure, so candidates such as Alberto Acosta, Rafael Correa and Norman Wray remain in the same position, meanwhile Guillermo Lasso and Mauricio Rodas have an initial different centers than in the no sentiment analysis case.
This indicates that in the period of time after they provided their answers in the survey till the time of this study, their opinions or actions have change or do not correspond at all with their answers provided during the survey.

In other words in the meantime, their perception with the voters has changed, the voters have a more realistic perception of the candidate.

Figure 37. Initial Candidates' Centers from Experts (Sentiment Analysis) Ecuador
7.4.2 Peru

For the case of Peru as in Ecuador the effect is repeated, it is possible to identify the differences among the location of the candidates. The effects can be observed between Figure 38 and Figure 39, where candidates such as Luis Castañeda remain with the same ideals and mentality as he answered in the survey.

![Figure 38. Initial Candidates' Centers (No Sentiment Analysis) Peru](image)

Candidates such as Kuczynski and Toledo have changed slightly in position and got closer to the voters, however Humala has changed his public image or ideas according to the voters and himself.

It is quite interesting to visualize such changes after the sentiment analysis consideration in the candidates profiles. It is hard to say that they have lied or it is part of an strategy that does not represent all his/her ideals. It seems that the candidates are always willing to change ideological positions in order to challenge or compete for the same voter’s segment with the closer opponent.
Figure 39. Initial Candidates’ Centers (Sentiment Analysis) Peru
7.5 Fuzzy Clustering Algorithm

After these comparisons between candidate profiles, it is possible to observe differences, but do these differences have a real impact in the vote decision of the users? It is time to compute the Fuzzy Clustering algorithm and find the differences.

7.5.1 Ecuador

In the case of Ecuador the Figure 40 provides the results of the final candidates’ clusters and these final clusters will determine later the degree of membership of each voter and the voter classification.

![Figure 40. Fuzzy Clustering Algorithm (No-Sentiment Analysis)](Ecuador)

In terms of comparison the differences between the Figure 40 and Figure 42. They do not offer a visual difference in terms of the calculation of the final centers. The final center positions remain equal and the sentiment changes do not have a big impact in the voters’ cluster classification. Later we will see with more details in the membership degree by voter in order to compare numerically, if there is an impact or a difference between both fuzzy clustering results for the case of Ecuador. Lastly, Figure 41 shows in colors the voters who belong to each political candidate, which is the same for both cases: sentiment and non sentiment analysis.
Figure 41. Fuzzy Clustering Algorithm Voters’ Clusters Ecuador

Figure 42. Fuzzy Clustering Algorithm (Sentiment Analysis) Ecuador
7.5.2 Peru

The case of Peru has a different behavior than the case of Ecuador, here the displacement between two candidates is not bigger as in the case of Ecuador but it has big consequences in terms of users cluster classifications.

The Figure 43 and Figure 45 show that there is a swap between two cluster centers, this a non expected result that has interesting consequences.

![Figure 43. Fuzzy Clustering Algorithm (No-Sentiment Analysis) Peru](image)

The final centers between Alejandro Toledo and Fujimori have swapped, so the voters between these two clusters have shifted the order as we can appreciate the change between the Figure 44 and Figure 46. Mathematically the voters that once belonged to one candidate belong to the other and vice versa but the interpretation of this could be that both candidates are always in constant rivalry and they try to give a similar public image and ideology to both voters groups in order to boost their popularity. This effort to play with their public image is causing uncertainty in the voters perception of the candidates. So voters may become unsure for which of the two candidates to vote.
Although these are some possible predictions to explain the behavior of the changes in data, it is important to remark that sentiment analysis allow us to fill this gap that was unknown before and grasp some understanding of what have happened after the survey took place.

Here the profile with sentiment analysis allows analysts, voters and candidates to see what they are doing wrong or good and be able to change it before it is too late to win an election. The importance and impact of social media these days is observed in the analysis of these datasets.
Figure 45. Fuzzy Clustering Algorithm (Sentiment Analysis) Peru

Figure 46. Fuzzy Clustering Algorithm Voters' Clusters (Sentiment Analysis) Peru
7.6 Degrees of Membership

It has been observed that in both election cases, there has been some influence from the social internet media and notice how the preferences of the candidates have changed in time but how much is much? and how much is few? The degrees of membership from the fuzzy cluster algorithm provides levels of membership among voters in order to grasp more details in numbers.

7.6.1 Ecuador

As it was mentioned before, the case of Ecuador did not bring too many changes in the final clusters in terms of voters classification but it shows a membership difference from a random voter (200) chosen in the dataset on purpose between two clusters.

Between Figure 47 and Figure 48, it is possible to compare the same voter between the scenario A and B and we observe that basically, the voter is between to vote between Mauricio Rodas and Guillermo Lasso. Although at the end it is assumed that he will vote for the candidate with major affinity (greater percentage). It is possible to recognize that Mauricio Rodas has increased his affinity with the voter rather than Alberto Acosta.
These small changes show that even if there is not a big distortion in the candidate’s ideology by sentiment analysis. The users are dynamic and tend to change their perception of the candidate.

There is an impact of what a candidate say and do also in their social networks, which opens many different questions for different studies in this field.
7.6.2 Peru

In contrast with Ecuador, Peru’s case suffered a big change in the distribution of two clusters and here is possible to observe how one user changed his ideology for the other candidate, in other words changed his vote to another candidate.

![Figure 49. Degrees of Membership (No-Sentiment Analysis) Peru](image)

The Figure 49 and Figure 50 show the difference between a voter (2) who is in the limit between two candidates and he/she is not be sure for whom to vote.

The voter originally at the time of the survey was wiling to vote for Keiko Fujimori but it turns out that as time goes by, the voter became more affine with Alejandro Toledo, so in theory if the elections will be today, at the time of the sentiment analysis was calculated, the user would have vote for Alejandro Toledo. Assuming that the voter (2) will vote for the candidate with closer affinity.
It remains interesting to observe the differences between two scenarios (A and B) and both are needed in order to find comparisons and good insights of the development of a political campaign or the perception of the voter about a certain candidate.

Unfortunately more research has to be done in this direction but as it was shown in this document, sentiment analysis in a candidate profile, plays an important role in the voter’s decisions and it remains as a future work to future researches to improve this model methodology.
8. Recommendations

The goal of this section is to provide recommendations for VAA Designers who aim to include sentiment analysis and Fuzzy Clustering Algorithms in their recommender systems. All recommendations are based on observations which have been made during the work on this master thesis.

- Sentiment analysis is a great element to shown the evolution of the perception of the users with respect some candidates, but it is important to have preferently a dataset with some predetermined social issue questions that can be answered by sentiment analysis also, otherwise it is possible to easily lose track of the study.

- In sentiment analysis, mapping tweets to social topics, can be improved by assigning a sentiment weighs based on tweets, rather than by words.

- In sentiment analysis there are several altmetrics that can be obtained, some of them were presented in this document such as number of followers or number of mentions in tags. However there are many more like counting the likes of a user or define geographical tweet radius, different language tweets, gender etc. The important is to define clearly what is the goal of the study to be performed with this data.

- Sentiment analysis can be useful if the social question topics are clear. It is hard in twitter to capture deep conversations with 140 characters limit. For instance avoid to generate complex questions such as do you agree that the article 3\textsuperscript{rd} of the constitution must be changed? People in twitter will mention the word constitution or the word change but it will be hard to find if these changes are related to the article 3\textsuperscript{rd} in an automatic way. Otherwise human intervention is needed which reduces the capacities of the automatic sentiment analysis.

- It is important to get a reliable sentiment analysis dictionary where you can compare the most often words and assign a weight to this word. In the time of this thesis was performed, there were many dictionaries for english language but nothing similar for Spanish language.
• Friends of Friends (FOAF) can be useful also to know the friends who are connected with the main profile, in order to analyses also the tweets of the acquaintances in the candidate social network.

• Before any sentiment analysis, it is pertinent to review if the candidates are active on the social networks, otherwise there will not be enough information to be compared. Use Euclidian distance first to grasp an idea of your dataset and understand which candidates offer most of the information and decide which ones are worth it to be monitored in social networks.

• Finally there is always place for innovation in the VAA’s algorithms that can be used to analyse the sentiment analysis data obtained. In the case of this thesis, the Fuzzy Clustering Algorithm has several benefits and it is very human readable when it is time to compute the membership function. The algorithm is not black or white, it has certain granularity that can be helpful to be interpreted. However there is a big domain to innovate with new VAA’s algorithms that can overcome the deficiencies of the current ones.
9. Conclusions

This study represents the first of its type to compare two different complex scenarios (with and without sentiment analysis) and it shows good insights that the sentiment analysis is an element nowadays that must be included as a part of any VAA’s study.

It also implements an innovative fuzzy clustering algorithm that presents rich information in the results, specially for classifying voters and their level of membership. Although this algorithm presents many advantages, it also has disadvantages such as the exclusion of some candidates because they did not have enough voters to balance the entire dataset, in order to have the same amount of voters regardless the popularity of the candidate. Fuzzy Clustering is a fair algorithm with a good level of details per voter but it can be biased if the dataset is not balanced or if there are more voters for a certain political party, then the algorithm will balance the centers of the final clusters, according to the most dominant of them.

During the elaboration of the thesis, many challenges were faced, specially for the programing part in python and matlab. Certain level of expertise was needed in order to obtain the correct information. Understanding the Twitter API was not trivial and it was needed some research to be done in order to obtain the tweets by a stream. In terms of MATLAB, it was needed to invest some time to understand and implement the Fuzzy Clustering Algorithm.

At the end, the thesis offered meaningful results and it was possible to answer all the research questions but it opened also new ones that can be part of future studies such as: What is the right level of granularity of my survey questions? How sentiment analysis can be more effective if more parameters are considered? Etc. Finally, this work provides an ideal base for future research in terms of Fuzzy Recommender Systems, Sentiment Analysis and may motivate to conduct further inquiries in the field of Voting Advice Applications.
10. Future Work

Last but not least, the sentiment analysis and the fuzzy clustering algorithm has shown good results and it has a potential to be tested in similar cases or current political elections. Although a Fuzzy Clustering analysis can be performed online, it is essential to build an automatic sentiment analysis tool that works 24h/7 days per week and this can be achieved with a simple Apache Spark opensource product as it is shown in the Figure 51.

![Figure 51. Real Time Sentiment Analysis Architecture](image)

The idea behind is to be able to monitor all the time the tweets because there might be a tweet from a third party that might have not been observed and it is relevant for the sentiment analysis. Therefore the figure 51 shows a potential architecture that uses SPARK to analyse the streams of the twits and make the sentiment analysis on the fly. Later the analysed data can be stored in a database and later a Fuzzy Recommender system or any other VAA application can gather the data and use it in their own system.

The development of such sentiment analysis monitoring can bring rich analysis and applications to the domain of VAAs and also help researchers to understand the impact of the social media on the voters.
11. References


12. Appendix

The aim of this section is to add the source code used during the analysis of the master thesis, please feel free to reuse the code and adapt it to your convenience. The programs were inspired by several opensource libraries and repositories and examples in https://dev.twitter.com/overview/api

12.1 Word Count Words Program

```python
# Author: Jose A. Mancera
# The program reads a textfile and count the words in the file
import sys

def word_count_dict(filename):
    word_count = {}  # Map each word to its count
    input_file = open(filename, 'r')
    for line in input_file:
        words = line.split()
        for word in words:
            word = word.lower()
            # Special case if we're seeing this word for the first time.
            if not word in word_count:
                word_count[word] = 1
            else:
                word_count[word] = word_count[word] + 1
    input_file.close()  # Not strictly required, but good form.
    return word_count

def print_words(filename):
    """Prints one per line '<word> <count>' sorted by word for the given file."""
    word_count = word_count_dict(filename)
    words = sorted(word_count.keys())
    for word in words:
        print word, word_count[word]

def get_count(word_count_tuple):
    """Returns the count from a dict word/count tuple -- used for custom sort."""
    return word_count_tuple[1]

def print_top(filename):
    """Prints the top count listing for the given file.
    word_count = word_count_dict(filename)
    items = sorted(word_count.items(), key=get_count, reverse=True)
    # Print the first 20
    for item in items[:20]:
        print item[0], item[1]

def main():
    option = '--topcount'
```


filename = 'FILENAME_TO_ANALYSE'
if option == 'count':
    print_words(filename)
elif option == 'topcount':
    print_top(filename)
else:
    print 'unknown option: ' + option
    sys.exit(1)

if __name__ == '__main__':
    main()

12.2 JSON To CVS Converter with UTF-8 Format

# Author: Jose A. Mancera
# The program creates a CVS file from JSON file, filtered by some parameters

import json
import csv
import io

data_json = io.open('INPUT_FILE.json', mode='r', encoding='utf-8').read()  # reads in the JSON file
data_python = json.loads(data_json)
csv_out = io.open('OUTPUT_FILE_utf8.csv', mode='w', encoding='utf-8')  # opens csv file

fields = u'created_at,text,screen_name,followers,friends,rt,fav'  # field names
csv_out.write(fields)
csv_out.write(u'
')

for line in data_python:
    # writes a row and gets the fields from the json object
    # screen_name and followers/friends are found on the second level hence two get methods
    row = [line.get('created_at'), '' + line.get('text').replace('"','"') + '"',
    # creates double quotes
    line.get('user').get('screen_name'),
    unicode(line.get('user').get('followers_count')),
    unicode(line.get('user').get('friends_count')),
    unicode(line.get('retweet_count')),
    unicode(line.get('favorite_count'))]

    row_joined = u','.join(row)
csv_out.write(row_joined)
csv_out.write(u'
')
csv_out.close()
12.3 Collect all the User Twitter Information From Users

#!/usr/bin/env python

""
Use Twitter API to grab user information from list of organizations; export text file

Uses Twython module to access Twitter API
""

import sys
import string
import simplejson
from tweython import Twython

#WE WILL USE THE VARIABLES DAY, MONTH, AND YEAR FOR OUR OUTPUT FILE NAME
import datetime
now = datetime.datetime.now()
day = int(now.day)
month = int(now.month)
year = int(now.year)

#FOR OAUTH AUTHENTICATION -- NEEDED TO ACCESS THE TWITTER API
t = Twython(app_key = 'CREDENTIALS',
    app_secret = ' CREDENTIALS ',
    oauth_token = ' CREDENTIALS ',
    oauth_token_secret = ' CREDENTIALS ')

# #REPLACE WITH YOUR LIST OF TWITTER USER IDS
ids = 'MashiRafael, LassoGuillermo, LucioGutierrez3, MauricioRodasEC, AbAlvaroNoboa, AlbertoAcostaE, normanwray, iglesiamds, Ollanta_HumalaT, KeikoFujimori, AlejandroToledo, ppkamigo, LuchoCastanedaL, '
users = t.lookup_user(screen_name = ids)

#NAME OUR OUTPUT FILE - %i WILL BE REPLACED BY CURRENT MONTH, DAY, AND YEAR
outfn = "twitter_user_data_%i.%i.%i.txt" % (now.month, now.day, now.year)

#NAMES FOR HEADER ROW IN OUTPUT FILE
fields = "id screen_name name created_at url followers_count friends_count statuses_count favourites_count listed_count contributors_enabled description protected location lang expanded_url".split()

#INITIALIZE OUTPUT FILE AND WRITE HEADER ROW
outfp = open(outfn, "w")
outfp.write(string.join(fields, "\t") + "\n") # header

for entry in users:
    #CREATE EMPTY DICTIONARY
    r = {}
    #ASSIGN VALUE OF 'ID' FIELD IN JSON TO 'ID' FIELD IN OUR DICTIONARY
    r['id'] = entry['id']
    #SAME WITH 'SCREEN_NAME' HERE, AND FOR REST OF THE VARIABLES
    r['screen_name'] = entry['screen_name']
    r['name'] = entry['name']
    r['created_at'] = entry['created_at']
    r['url'] = entry['url']
    r['followers_count'] = entry['followers_count']
    r['friends_count'] = entry['friends_count']
    r['statuses_count'] = entry['statuses_count']
    r['favourites_count'] = entry['favourites_count']
    r['listed_count'] = entry['listed_count']
    r['contributors_enabled'] = entry['contributors_enabled']
    r['description'] = entry['description']
```
r['protected'] = entry['protected']
r['location'] = entry['location']
r['lang'] = entry['lang']

# NOT EVERY ID WILL HAVE A 'URL' KEY, SO CHECK FOR ITS EXISTENCE WITH IF CLAUSE
if 'url' in entry['entities']:
    r['expanded_url'] = entry['entities']['url']['urls'][0]['expanded_url']
else:
    r['expanded_url'] = ''

print r

# CREATE EMPTY LIST
lst = []

# ADD DATA FOR EACH VARIABLE
for f in fields:
    lst.append(unicode(r[f]).replace("\", "/"))

# WRITE ROW WITH DATA IN LIST
outfp.write(string.join(lst, "\t").encode("utf-8") + "\n")
```

```
12.4 Tweets Capture for a User Account (3200 Tweets)

#!/usr/bin/env python
# encoding: utf-8
import sys
import tweepy
#https://github.com/tweepy/tweepy
import csv

# Twitter API credentials
consumer_key = 'CREDENTIALS'
consumer_secret = 'CREDENTIALS'
access_key = 'CREDENTIALS'
access_secret = 'CREDENTIALS'

def get_all_tweets(screen_name):
    # Twitter only allows access to a user's most recent 3240 tweets with this method

    # authorize twitter, initialize tweepy
    auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_key, access_secret)
    api = tweepy.API(auth)

    # initialize a list to hold all the tweepy Tweets
    alltweets = []

    # make initial request for most recent tweets (200 is the maximum allowed count)
    new_tweets = api.user_timeline(screen_name = screen_name, count=200)

    # save most recent tweets
    alltweets.extend(new_tweets)

    # save the id of the oldest tweet less one
    oldest = alltweets[-1].id - 1

    # keep grabbing tweets until there are no tweets left to grab
    while len(new_tweets) > 0:
        print "getting tweets before %s" % (oldest)

        # all subsequent requests use the max_id param to prevent duplicates
        new_tweets = api.user_timeline(screen_name = screen_name, count=200, max_id=oldest)

        # save most recent tweets
        alltweets.extend(new_tweets)

        # update the id of the oldest tweet less one
```

```
oldest = alltweets[-1].id - 1

print "...%s tweets downloaded so far" % (len(alltweets))

#transform the tweepy tweets into a 2D array that will populate the csv
outtweets = [[tweet.id_str, tweet.created_at, tweet.text.encode("utf-8")]
for tweet in alltweets]

#write the csv
with open('%s_tweets.csv' % screen_name, 'wb') as f:
    writer = csv.writer(f)
    writer.writerow(['id', 'created_at', 'text'])
    writer.writerows(outtweets)

pass

if __name__ == '__main__':
    # pass in the username of the account you want to download
get_all_tweets(sys.argv[1])

12.5 Twitter Stream Listener For Many Accounts

'''
Twitter Stream Listener
'''

# tweepy setup
import time
from tweepy import Stream
from tweepy import OAuthHandler
from tweepy.streaming import StreamListener
import io
import os
import json

# twitter OAuth
ckey = ""
consumer_secret = ""
access_token_key = ""
access_token_secret = ""

# Listener Class Override
class listener(StreamListener):
    def __init__(self, start_time, time_limit=6000):
        self.time = start_time
        self.limit = time_limit
        self.tweet_data = []

    def on_data(self, data):
        saveFile = io.open('raw_tweets_Ecuador.json', 'a', encoding='utf-8')
        while (time.time() - self.time) < self.limit:
            print (time.time() - self.time, 'hasta ', self.limit)
            try:
                self.tweet_data.append(data)
                return True
            except BaseException, e:
print 'failed ondata,', str(e)
time.sleep(10)

pass

saveFile = io.open('raw_tweets_Ecuador.json', 'w', encoding='utf-8')
saveFile.write(u'
')
saveFile.write(''.join(self.tweet_data))
saveFile.write(u'
')
saveFile.close()
exit()

def on_error(self, status):
    print status

def on_disconnect(self, notice):
    print 'bye'

# Beginning of the specific code
start_time = time.time()  # grabs the system time

keyword_list = ['MashiRafael', 'LassoGuillermo', 'LucioGutierrez3',
'MauricioRodasEC', 'AbAlvaroNoboa', 'AlbertoAcostaE', 'normanwray',
'iglesiamds']  # track list

auth = OAuthHandler(ckey, consumer_secret)  # OAuth object
auth.set_access_token(access_token_key, access_token_secret)

twitterStream = Stream(auth, listener(start_time,
time_limit=36000))  # initialize Stream object with a time out limit
twitterStream.filter(track=keyword_list,
languages=['es'])  # call the filter method to run the Stream Listener

12.6 Sentiment Analysis Dictionary Comparison (Code Section)

def print_happiest_words(rdd):
    top_list = rdd.take(5)
    for tuple in top_list:
        print("%s (%d)" % (tuple[1], tuple[0]))

if __name__ == "__main__":

    # Read in the word-sentiment list and create a static RDD from it
    word_sentiments_file_path = "AFINN-Spanish.txt"
    word_sentiments = textFile(word_sentiments_file_path) \
    .map(lambda line: tuple(line.split("\t")))

    lines = open(sys.argv[1])

    word_counts = lines.flatMap(lambda line: line.split(" ") \
    .map(lambda word: (word, 1)) \
    .reduceByKey(lambda a, b: a + b)

    # Determine the words with the highest sentiment
    # with the static RDD inside the transform() method and then multiplying
    # the frequency of the words by its sentiment value
    happiest_words = word_counts.transform(lambda rdd: word_sentiments.join(rdd)) \
    .map(lambda (word, tuple): (word, float(tuple[0]) * tuple[1])) \
    .map(lambda (word, happiness): (happiness, word)) \
    .transform(lambda rdd: rdd.sortByKey(False))

    happiest_words.foreachRDD(print_happiest_words)