A dissertation submitted to the University of Greenwich in partial fulfilment of the requirements for the Degree of Master of Science in Data Warehousing & Business Intelligence

Twitter Opinion Mining and Visualization of Ecuadorian Government’s Decisions

Name: Lina Saquicela

Student ID: 000793716

Supervisor: Dr Simon Scola
Submission Date: 18 September 2014
Word count: 15090
TWITTER OPINION MINING AND VISUALIZATION OF ECUADORIAN GOVERNMENT’S DECISIONS

Computing & Mathematical Sciences, University of Greenwich, 30 Park Row, Greenwich, UK.

(Submitted 18 September 2014)

ABSTRACT. Social media impact on the present time is undeniable. People communication dynamics are extremely different from those that were available a decade or fifteen years ago. These new age interactions have allowed people expressing their ideas or opinions in an extended way. On the other hand, such interactions have encouraged social actors to take into consideration these opinions in order to accomplish different goals.

Thus, the analysis of data generated by people has become a popular task that is required for different applications. One application can be found at political or government spheres. Specifically, sentiment analysis applications have been implemented in order to evaluate the political scenario before elections or to improve e-government initiatives.

Following such approach, the present project introduces the opinion mining of selected Ecuadorian Government’s decisions data generated on Twitter, and the presentation of such results through a Web application. This way, a hybrid sentiment analysis method has been implemented and utilised to classify by its polarity the tweets corresponding to three decisions taken by the Ecuadorian government, which have caused a certain degree of discussion on Twitter.

Manual classifications of a portion of the tweet sets have been carried out as well as the utilisation of an online sentiment analysis API for classifying all the tweets and comparing the obtained results.

All these implementations and the respective development process allowed evidencing the challenge and limitations of this type of solutions. Although the hybrid method implementation outperformed the sentiment analysis API on some datasets, it was not good enough for all the datasets. In addition the method’s results are clearly far from satisfactory. However the obtained solution can be considered a good starting point that open the doors to future improvements in order to achieve a solid tool.

Keywords: Data Mining, Twitter, Government, Information Visualization, Business Intelligence, Social media, Big Data, Sentiment Analysis.
Acknowledgements

I would like to thank Dr Simon Scola for agreeing to be my supervisor and for his feedback and guidance, throughout the process, starting from the proposal delivery to the conclusion of this MSc project.

I want to thank both Dr Simon Scola and Dr Tatiana Simmonds for agreeing to have the project demonstration after the submission of this document.

I would also like to thank Dr Ronan Cummins, for promoting the interest on such topics throughout the Business Intelligence and Data Mining course as well as the Big Data course, imparted during the MSc programme.
# Table of Contents

Abstract .......................................................................................................................... i
Acknowledgements ......................................................................................................... ii

Table of Contents .......................................................................................................... iv
List of Figures .................................................................................................................. vii
List of Tables .................................................................................................................. ix
List of Acronyms ........................................................................................................... x

1 Introduction .................................................................................................................... 1
  1.1 Overview .................................................................................................................. 1
  1.2 Report Roadmap ..................................................................................................... 3
2 Literature Review ......................................................................................................... 5
  2.1 Sentiment Analysis ............................................................................................... 5
    2.1.1 Definition ........................................................................................................ 5
    2.1.2 Sentiment Analysis relevance ....................................................................... 5
    2.1.3 Opinion and Twitter ....................................................................................... 6
    2.1.4 Sentiment Analysis Process .......................................................................... 6
    2.1.5 Analysis levels ............................................................................................... 6
    2.1.6 Methods ......................................................................................................... 7
    2.1.7 Sentiment Analysis Challenges .................................................................... 8
  2.2 Related Work ......................................................................................................... 10
    2.2.1 Areas .............................................................................................................. 10
    2.2.2 Related applications ....................................................................................... 10
  2.3 Summary ............................................................................................................... 12
3 Project Initiation .......................................................................................................... 13
  3.1 Legal, Social, Ethical and Professional issues ....................................................... 13
    3.1.1 Legal Issues ................................................................................................... 13
    3.1.2 Social Issues ................................................................................................. 13
    3.1.3 Professional Issues ....................................................................................... 14
    3.1.4 Ethical Issues ............................................................................................... 14
3.2 Research Method ............................................................................................................. 15
  3.2.1 Research Framework .................................................................................................. 15
  3.2.2 Motivation ................................................................................................................ 15
  3.2.3 Research questions ..................................................................................................... 15
  3.2.4 Research plan ............................................................................................................ 15
3.3 Government’s Decisions ................................................................................................. 16
  3.3.1 Ecuadorian Government Background ....................................................................... 16
  3.3.2 Selection of Government’s Decisions ..................................................................... 17
3.4 Selection of Methods and Tools ....................................................................................... 18
  3.4.1 Sentiment Analysis Method .................................................................................... 18
  3.4.2 Tweets extraction ...................................................................................................... 19
  3.4.3 Storage ....................................................................................................................... 19
  3.4.4 Data Mining Tools and Visualisation ...................................................................... 20
3.5 Development Methodology ............................................................................................. 20
  3.5.1 Justification ............................................................................................................... 21
  3.5.2 Techniques .............................................................................................................. 21
  3.5.3 Lifecycle Model ........................................................................................................ 21
3.6 Summary .......................................................................................................................... 22
4 Analysis and Design ............................................................................................................ 23
  4.1 Analysis Task ................................................................................................................. 23
    4.1.1 Assumptions Definition ......................................................................................... 23
    4.1.2 Requirements List ................................................................................................. 23
    4.1.3 Requirements Prioritisation ................................................................................. 24
    4.1.4 Iteration Plan ........................................................................................................ 25
    4.1.5 Proposed solution ................................................................................................. 25
  4.2 Web Application Analysis ................................................................................................ 26
    4.2.1 Web application overview .................................................................................... 26
    4.2.2 User / Audience analysis ...................................................................................... 26
    4.2.3 Workflow analysis ................................................................................................. 29
  4.3 Design ............................................................................................................................. 30
    4.3.1 Web application pages .......................................................................................... 30
    4.3.2 Database ................................................................................................................ 32
    4.3.3 Font for Web pages ............................................................................................... 33
  4.4 Summary .......................................................................................................................... 34
5 Implementation

5.1 First iteration

5.1.1 Development environment setup

5.1.2 Twitter data extraction and storage

5.1.3 Manual sentiment classification

5.1.4 Sentiment Analysis module

5.1.5 Data Visualisation module

5.1.6 Assessment module

5.2 Second iteration

5.2.1 Sentiment Analysis module

5.2.2 Data Visualisations module

5.3 Third iteration

5.3.1 Sentiment Analysis module

5.3.2 Data Visualisations module

5.4 Summary

6 Results and Evaluation

6.1 Results

6.1.1 Julian Assange Asylum decision

6.1.2 Yasuni exploitation decision

6.1.3 2011 Referendum

6.2 Evaluation

7 Conclusions and Future Work

8 References
List of Figures

Figure 1: Solution general perspective ................................................................. 25
Figure 2: Solution General View in the iterative process ........................................ 26
Figure 3: Persona Silvia ....................................................................................... 28
Figure 4: Persona Fernando ................................................................................ 28
Figure 5: Persona Eva ......................................................................................... 29
Figure 6: Main interactions ................................................................................ 29
Figure 7: Home page ......................................................................................... 30
Figure 8: Opinion page ..................................................................................... 31
Figure 9: Methods Assessment page .................................................................. 32
Figure 10: Document-oriented data model for Twitter data .................................. 32
Figure 11: Results Entity-Relationship model .................................................... 33
Figure 12: Relational Database tables ............................................................... 33
Figure 13: Development environment ................................................................ 35
Figure 14: Tweets extraction and storage ............................................................ 37
Figure 15: Application for Manual Polarity Classification ................................... 37
Figure 16: Twitter data manual classification .................................................... 38
Figure 17: URL’s removal log ............................................................................ 38
Figure 18: Pre-processing module ..................................................................... 39
Figure 19: sa_box_client application log ............................................................ 40
Figure 20: Use of Sentiment analysis tool ........................................................ 40
Figure 21: Home page ....................................................................................... 41
Figure 22: Opinion results for manual classification ......................................... 42
Figure 23: Opinion results for classification using “Apicultur” API .................... 42
Figure 24: Number of tweets per day ................................................................ 43
Figure 25: Apicultur method assessment results ................................................. 43
Figure 26: Confusion matrix .......................................................................... 44
Figure 27: Hybrid method implementation modules ........................................... 45
Figure 28: Pre-processing execution log ............................................................. 46
Figure 29: Sentiment feature generator process ............................................... 46
Figure 30: Tagging process ........................................................................... 47
Figure 31: Sentiment Features generation results .............................................. 48
Figure 32: Sentiment Analysis results using Hybrid method ................................ 49
Figure 33: Hybrid Method assessment results ................................................ 49
Figure 34: Tag results after introduced modifications ......................................... 50
Figure 35: Sentiment Analysis results using Modified Hybrid method .............. 50
Figure 36: Modified Hybrid method assessment results .................................... 51
Figure 37: Assange decision - Number of tweets per day .................................. 52
Figure 38: Assange decision - Manual Classification ....................................... 52
Figure 39: Assange decision – Apicultur API analysis ....................................... 53
Figure 40: Assange decision - Hybrid method analysis ...................................... 53
Figure 41: Assange decision - Modified Hybrid method analysis ..................... 53
Figure 42: Assange decision - Methods assessment .......................................... 54
Figure 43: Yasuni decision - Number of tweets per day .................................... 54
Figure 44: Yasuni decision - Manual Classification ........................................... 55
Figure 45: Yasuni decision – Apicultur API analysis .................................................. 55
Figure 46: Yasuni decision - Hybrid method analysis .................................................. 55
Figure 47: Yasuni decision - Modified Hybrid method analysis................................. 56
Figure 48: Yasuni decision - Methods assessment...................................................... 56
Figure 49: 2011 Referendum decision - Number of tweets per day............................. 57
Figure 50: 2011 Referendum decisions - Manual Classification .................................... 57
Figure 51: 2011 Referendum decisions – Apicultur API analysis ................................. 57
Figure 52: 2011 Referendum decisions - Hybrid method analysis ............................... 58
Figure 53: 2011 Referendum decisions - Modified Hybrid method analysis ............... 58
Figure 54: 2011 Referendum decisions - Methods assessment ...................................... 59
List of Tables

Table 1: Research plan ............................................................................................................. 16
Table 2: Ecuadorian Government's Decisions ......................................................................... 18
Table 3: Prioritised Requirements .......................................................................................... 24
Table 4: Iteration Plan ............................................................................................................. 25
Table 5: Tweets extracted summary ....................................................................................... 36
Table 6: Tables in relational database .................................................................................... 41
List of Acronyms

API – Application Programming Interface
DSDM - Dynamic Systems Development Method
JSF – JavaServer Faces
JSON – JavaScript Object Notation
NoSQL – Not only SQL
SVM – Support Vector Machines
1 Introduction

1.1 Overview

The proposed project aims to perform sentiment analysis on particular Twitter datasets and present the results via a Web application. The Twitter datasets consist of tweets related to three relevant decisions taken by the Ecuadorian Government. The purpose of this project is to provide an initial development on this area that can potentially become a useful tool for the Ecuadorian Government Communication personnel or any interested communication agency.

The importance of the study of opinion mining or sentiment analysis field can be evidenced in the increasing quantity of research works and literature resources. This growing interest has allowed obtaining significant improvements in the accuracy of the results, performance-wise, as well as a consolidated commercial offering of APIs and tools, product-wise.

However, the majority of authors are of the opinion that the study is far to be satisfactorily completed. This is why, from an academic context, it is beneficial to carry out implementations working over different domains and circumstances as this may contribute to obtain new improvements and perspectives.

On one hand, the current importance of social media as well as the quantity and quality of generated data encourage businesses and organisations to utilise this resource for diverse purposes. On the other hand, the politics, media and social media scenario is very arguably; their influence, intentions, roles and how they affect each other are complex and controversial topics.

In the case of Ecuadorian Government, social media is considered a great alternative for communication purposes. Although excesses can arise, the effectiveness of social media is highly recognized.

Therefore, the government decided to develop an official communication platform that consists of traditional media as television, radio and newspaper, as well as social media presence. For which, official accounts were created in the most popular social media sites such as: Twitter, Facebook, Flickr, YouTube and Ivoox. This platform has been consolidated since the government confirmed that citizens have been misinformed several times, which generated unjustified issues.

As a result, it is observable that authorities and offices use actively such platform and citizens are encouraged to participate; this initiative, as in the rest of world, includes the active participation of the president and vice-president in social media, who in fact can be contacted through their personal accounts.

Thus, the use of social media by political actors has motivated the interest in developing tools that may support legitimate purposes and obligations such as the adequately information delivery and the political stability protection of the nation.
Consequently, the aspects that motivate the chosen project are, on one hand, the interest in Government Intelligence topics, and on the other hand, the opportunity of delving into complex and novel topics. Such aspects will help to develop new skills and acquire knowledge in tools and techniques that can be utilised at the organisational level for the adoption of new technologies. In addition, the involvement in such type of implementations may open the doors for new job opportunities.

Therefore, the proposed project pretends to cover these important topics and utilise the available resources in order to explore areas as Big Data Analytics, Data Mining and Data Visualisations, as well as participate in the development process of cutting-edge Business Intelligence solutions.

Below are presented the objectives proposed for this project:

1. To research and select appropriate tools and techniques for implementation
   The achievement of this objective requires a literature review in order to determine the techniques and algorithms that will be used to perform the sentiment analysis as well as the tools that will be used for the implementation.

2. To collect and store the required tweets
   This objective requires the selection of significant Ecuadorian Government’s decisions for later Twitter data extraction. In addition, it is also necessary to select a suitable database to store the extracted data.

3. To perform a sentiment analysis
   For achieving this objective it is necessary to follow a data mining process that will include the use of several techniques to pre-process the unstructured data and different algorithms that allow categorizing the tweets by the expressed opinion.

4. To design and implement a Web application for presenting results
   The accomplishment of this objective requires following a user centred design process that includes the definition and analysis of requirements, design, implementation, deployment and testing, all tasks following an agile development methodology.

5. To evaluate the analysis outcome and final product
   This will be achieved by analysing and discussing the sentiment analysis results and assessing the products produced.

For the execution of this project it has been planned to follow the engineering research method, the DSDM Agile development methodology and a User-centred design process.

In order to utilise the appropriate elements, various selection processes have been planned. Thus, the government’s decisions, data storage, sentiment analysis method, programming language and visualisation libraries were selected after evaluating several alternatives.

For the sentiment analysis task, the method’s performance has been established as an essential factor to be taken into consideration. By selecting a sentiment analysis method that has shown good results, it is attempted to establish a good starting point for proposing modifications and new initiatives that can improve the current performance and eventually
offer a contribution to the knowledge. Thus, the limitations and challenges that a new scenario may impose will help to generate improvement suggestions.

For evaluation purposes it is planned to utilise an out-of-the-box sentiment analysis tool as well as perform a manual classification on some data. This way, it will be possible to determine the accuracy of the sentiment analysis methods utilised and compare their performances.

1.2 Report Roadmap

The present document aims to describe the process followed in order to accomplish the project’s objectives. After exposing the theoretical foundations, each section details the tasks that have been performed and the results obtained. Finally, the last sections discuss these results for concluding the document.

The following chapter presents the literature review elaborated for this project, in which the required theoretical foundation is summarised. The main Sentiment analysis topics discussed in this chapter are: definition and importance, analysis process and analysis levels, methods and challenges. A brief discussion on the opinion shared through Twitter is also included. Ultimately, a condensed review of related research works and applications has been appended, too.

The third chapter, firstly include the social, legal, ethical and professional considerations. Later, the initial project’s tasks are described. These tasks comprise the selection of the research method and development methodology, the selection of the Government’s decisions and the selection of the appropriate methods and tools to be utilised. For each selection process, the reasons that led to choose an alternative are presented and discussed opportunely.

The fourth chapter presents the Analysis and Design tasks. The Analysis comprises the Requirements Definition and Prioritisation as well as the Iteration Plan based on such requirements, as suggested by the selected development methodology. Later, it is presented the application of user-centred analysis techniques such as user/audience analysis and persona representations. Finally the chapter presents the design considerations and decisions made at database and user interface levels. Such ideas are presented through mock-ups and appropriate schemas. This section also includes a discussion of the font type and sizes arranged for the proposed Web pages.

The fifth chapter gives a detailed explanation on how the solution proposed has been implemented. The implementation is described according the planned iterations. This way, the initial steps as the development environment setup and data storage are presented. The processes and products corresponding to the Sentiment Analysis and Data Visualisation modules are introduced incrementally, following the marked iterations.

The sixth chapter presents the results obtained after completing the implementation task. These results comprise the assessment values obtained by performing calculations as well
as the products, applications, etc. that have been developed. All the results are finally explored and a critical discussion on what has been obtained is also presented.

The final chapter is dedicated to present the conclusions generated throughout the entire process and especially after obtaining the sentiment analysis results. Then, this section concludes by presenting further work suggestions.
2 Literature Review

2.1 Sentiment Analysis

2.1.1 Definition

Sentiment Analysis is a field of study of current interest. Many research works were conducted during the past decade that have achieved significant progress in this area, however it is a field that still requires more research efforts to achieve all its goals satisfactorily.

Sentiment Analysis attempts to extract the opinions in a text and find its associated polarity. Opinions express sentiment orientations towards targets that can be positive, negative or even neutral. An opinion can be found in subjective expressions as well as in objective expressions. In addition, opinions can be stated in regular sentences or comparative sentences; and they can be expressed in text, explicitly or implicitly (Liu, 2012). The variety of opinion types makes them difficult to examine and some of them are not considered in order to simplify their study.

The term “Sentiment Analysis” has many equivalent terms, among which, the term “Opinion Mining” is used frequently. Although many authors use these terms interchangeably, (Cambria & Hussain, 2012) make a differentiation between these terms referring to polarity detection as Opinion mining and emotion recognition as Sentiment analysis. Even so, they also explain that the identification of emotional states is usually part of polarity detection and therefore both terms indicate the same field of study.

Thus, diverse emotion typologies and classifications have been utilised in order to detect polarity due to its strong connection to attitudes and feelings. For instance, (Cambria & Hussain, 2012) propose a novel model called Hourglass that organises primary emotions in dimensions and levels that can be combined in order to obtain a more complete range of emotional states. Then, then model is used for polarity detection purposes.

Although emotions can be utilised for detecting sentiment, opinions are not equivalent to emotions, as opinions do not necessarily express emotions and many emotions do not express any opinion (Liu, 2012).

2.1.2 Sentiment Analysis relevance

The relevance of Sentiment Analysis relies on the importance that opinion has these days. People have the need to express their opinions and, on the other side, individuals and organisations are interested in evaluating others’ opinion with diverse purposes. For this, it is necessary to examine texts and determine the sentiment on something.

Thus, leveraging the vast volume of opinionated data available by virtue of social media boom, many organisations of almost every domain have opted to implement sentiment analysis applications (Liu, 2012). Precisely, the immense amount of data makes it essential the use of automated systems to identify, extract and summarise people’s opinions.
And depending on how the analysis is done, the results may highlight the important trends or may lead to wrong conclusions (Zadrozny & Kodali, 2013). Hence the interest in studying this subject continues with the goal of overcoming its current limitations.

2.1.3 Opinion and Twitter

Social media collect vast amounts of opinionated posts from people every day, which reveals the significant role that social media perform around the world and the inevitable influence that has had on worldwide events (Liu, 2012). Hence, the interest in studying this kind of data has grown over the past few years.

Data from Twitter is used very often due to its ease of obtaining and unique characteristics. The site’s mechanical, text length limit and users’ characteristics make Twitter data different from other blogs or review sites. Although some of these characteristics may simplify the sentiment extraction, other characteristics impose unique challenges; therefore its study can be really enriching (Danneman & Heimann, 2014).

2.1.4 Sentiment Analysis Process

In general, the Sentiment Analysis process comprises two main tasks: retrieval of contents related to a topic of interest and polarity measurement of each item of data (Danneman & Heimann, 2014). Still, there are other tasks that can also be performed as part of this process as spam detection and quality determination, which seek to ensure the validity of the data to be analysed (Liu, 2012).

Some research works as the one conducted in (Khan, et al., 2014) presented the inclusion of pre-processing steps in order to overcome certain limitations and achieve a higher accuracy. The pre-processing task is executed before sending Twitter data to the sentiment classifier, and includes steps as replacement of abbreviations, lemmatization and spell checking.

Depending on the selected method, the polarity measurement task may include sub-tasks as lexicon generation or training data obtainment.

2.1.5 Analysis levels

Polarity detection has been performed at various levels such as document, sentence, phrase, aspect and word levels. However, the following three levels have been mainly studied (Liu, 2012).

2.1.5.1 Document level

A whole opinion document is classified as having a positive or negative sentiment. For this, it is assumed that the document contains opinions on a single entity, expressed by a single opinion holder.
2.1.5.2 **Sentence level**

A sentence is classified as having a positive or negative sentiment or not having opinion at all. For this, it is assumed that the sentence contains a single opinion from one opinion holder. This level of analysis comprises two sub-tasks: subjectivity classification and sentiment classification.

In this context, Twitter data can be analysed at sentence level. Although, a tweet may contain more than one sentence, its text length is restricted to 140 characters or 11 words approx., which is totally opposed to other social media data like blogs or product review sites that must be analysed using a document level approach (Danneman & Heimann, 2014).

2.1.5.3 **Aspect level**

In this case, the analysis is performed on what is said about a specific entity or target. Thus, the analysis does not seek for language constructs but rather for individual opinions.

This analysis comprises two main sub-tasks: aspect extraction and aspect sentiment classification. Depending on the method required, sometimes it is assumed that the targets are known.

2.1.6 **Methods**

In general, opinion mining research works have followed two main approaches for processing texts: Natural Language Processing and Semantic Web Approach. The former uses implicit representation of meaning to define similarities between texts while the latter uses semantic annotations to map text to the domain ontology (Sobkowicz, et al., 2012).

A more specific approach categorisation proposes the following categories (Cambria, et al., 2013):

- Keyword spotting that classifies text by detecting the presence of affect words.
- Lexical affinity that, in addition to keyword spotting, evaluates arbitrary words by assigning them a probability of indicating a particular affect.
- Statistical methods that utilise machine learning algorithms to learn the affective valence of words by using training data previously labelled.
- Concept-based approaches that utilise Web ontologies to analyse text considering implicit meanings associated with natural language concepts.

The methods utilised by the first and second approaches correspond to a lexicon-based sentiment classification. Such methods require a collection of words or phrases, each one associated to a polarity score (Danneman & Heimann, 2014).

The statistical methods utilised by the third approach correspond to machine learning classifiers. In turn, these classifiers can utilise supervised or unsupervised learning algorithms. Some popular supervised methods include the Naive Bayes Classifier,
Maximum Entropy and Support Vector Machines (SVM). On the other hand, unsupervised learning includes techniques such as Item Response Theory (IRT) and syntactic patterns.

The concept-based approaches include a novel approach called Sentic Computing which makes use of AI and Semantic Web techniques as well as common sense knowledge bases to develop intelligent engines (Cambria & Hussain, 2012).

In order to overcome some well-known limitations, other approaches and techniques have been exploited.

As it has been demonstrated that the sentiment classification performance is highly domain dependent, several cross-domain solutions have been proposed. These solutions include the use of machine learning algorithms to transfer the sentiment classifier to new domains, as well as the use of methods, such as Structural Corresponding Learning (SCL), for domain adaptation (Liu, 2012).

Other difficult that motivates researches to propose alternatives is that the majority of research work has been done in English and consequently such resources and tools are not available in other languages. Therefore, it has been proposed some cross-language sentiment classification solutions. One approach proposed is the use of translators to obtain the English version of the data, or to translate the resources to the required languages. Other approaches propose the use of methods such as Structural Correspondence Learning and Topic Modelling to extract features or topics shared across languages (Liu, 2012).

**Lexicon generation**

In the case of using a lexicon-based approach, it is necessary to have the list of sentiment words, phrases and idioms, or also called sentiment lexicon. In order to compile the lexicon, it has been proposed the following approaches (Liu, 2012):

- Dictionary-Based approach that uses dictionaries to obtain synonyms and antonyms for some seed sentiment words, iteratively, until no more words can be found. Other alternative proposed is to use a distance-based method for sentiment words assessment, starting from a given adjective.
- Corpus-based approach that uses a domain corpus for two different tasks. First, to discover other sentiment words starting from a general-purpose lexicon and second, to adapt a general-purpose lexicon to a new domain.

**2.1.7 Sentiment Analysis Challenges**

Below is presented a list of sentiment analysis challenges and limitations mentioned by several authors. Such problems have motivated the study and development of new and improved approaches throughout the years.

- Domain and context dependency
  Sentiment words may have different polarities depending on the domain where they are used. Even, when they are used in the same domain, some sentiment words
depend on the context of the containing sentence (Liu, 2012). In addition, phrases or expressions can also be domain and context dependents (Cambria & Hussain, 2012).

- **Language problem**
  As English texts have been the focus of the majority of research works, the majority of resources are available only in English (Cambria & Hussain, 2012).

- **Sentiment expression**
  Not all sentences that contain sentiment words express sentiment necessarily. This may happen in interrogative or conditional type of sentences. On the other hand, some sentences with no sentiment words express facts but can also imply opinions (Liu, 2012).

- **Sarcasm**
  Sarcastic sentences, having or not sentiment words are very common in certain areas as politics (Liu, 2012).

- **NLP problems**
  Authors also mention the significant problems that are still unresolved topics in Natural Language Processing. These are listed below:
    - Co-reference resolution that tries to determine which expressions refer to the same thing; and Word sense disambiguation that tries to determine if a word instance has been used subjectively or objectively (Liu, 2012).
    - Negation that can reverse meaning; and semantic weakness in statistical methods (Cambria & Hussain, 2012).
    - Use of adjectives and superlatives to express intensification (Danneman & Heimann, 2014).

- **Opinion Spam**
  It refers to the publication of fake opinions through the Web in order to satisfy personal or business interests. Therefore the detection of spamming is a necessary task to ensure the validity of the collected data. Although some techniques have been proposed to analyse the posts as well as the user’s behaviours, it is still considered a difficult task even manually (Liu, 2012).

- **Concept-based approach limitations**
  Although a novel proposal as the Concept-based approach may overcome several limitations, it still has to consider the following (Cambria & Hussain, 2012):
    - The approach is dependent on the knowledge bases richness and requires an adequate exploitation of them.
    - The approach is still context dependent and lacks of time representation for the analysis.

- **Twitter data challenges**
  Due to Twitter characteristics, posted sentences may be incomplete or poorly expressed. They may have bad grammar, poor punctuation, wrong spelling or the inclusion of words spelled differently in order to exaggerate. In addition, data sparsity may also be a problem.
The solutions that have been proposed are not able to consider all the possible cases given the complexities of each domain and context. Therefore, all current approaches have their own drawbacks and there is still much work to do in order to improve them.

2.2 Related Work

2.2.1 Areas

Politics is considered a very challenging area that may be really fruitful for sentiment analysis. As a result, many applications that analyse political discussions have been proposed.

Some applications, for instance, try to predict the results of an election or at least identify the issues and positions that matter to voters. The 2008 and 2012 American presidential elections were two major events that allow the implementation of such applications (Sharda & al, 2014).

Other applications perform sentiment analysis of data corresponding to specific political events in order to capture patterns. For instance, the uprisings in the Arab world have activated the study of social media data due to its evident role in such events (Younus, et al., 2011).

Government intelligence is other application scenario for sentiment analysis. In this case, government agencies may be interested in monitoring people’s opinions (Sharda & al, 2014). Such type of applications can benefit policymaking in governments, by providing them necessary feedback to set required actions, formulate policies and evaluate policy impact (Sobkowicz, et al., 2012).

2.2.2 Related applications

In (Khan, et al., 2014) is proposed a framework to classify tweets by its polarity, that includes a pre-processing module and a hybrid sentiment classifier. The pre-processing tasks comprise the removal of duplicated tweets, removal of URLs, hashtags and usernames; replacement of abbreviations and slangs, spell correction, removal of stop words, lemmatization and removal of special characters, except emoticons.

The hybrid sentiment classifier utilises three different classification techniques one after another, to finally return tweets labelled as positive, negative or neutral. The mentioned techniques are: Enhanced Emoticon Classifier (EEC), Improved Polarity Classifier (IPC) and SentiWordNet Classifier (SWNC). The classifiers use a set of emoticons, a list of positive and negative words and the SentiWordNet dictionary, respectively, in order to determine a tweet polarity.

The dataset utilised consists of 2116 tweets that were classified with an average accuracy of 85.7%, precision of 85.3% and recall of 82.2%. Thus, it was concluded that the proposed algorithm performs better when compared with similar work. Finally, it was suggested the
use of supervised learning algorithms and the comparison with other applications such as TweetFeel and Sentiment140.

In (Mejova, et al., 2013) is presented a general-purpose political sentiment classification process that uses the SVMlight classification tool combined with a logistic regression classifier. The classifiers are used to track the sentiment on the time span around US candidate debates in 2011. This is done with the goal of predicting changes of sentiment due to debates.

The dataset utilised consists of a sample of 10000 documents for each debate, and the overall accuracy obtained is of 54.4%. In this case, after comparing the results to national polls, it was concluded that the sentiment determine by the classification did not correspond well to the sentiment shown in polls.

The authors in (Younus, et al., 2011) proposed a method for sentiment classification that takes into account the users’ interactions and engagement in conversations. The method comprises the use of selected social features to gather the dataset and applying a Naive Bayes Classifier for polarity detection.

The dataset consisted of tweets collected using the query term ‘Tunisia’ for a period of 13 days that were classified with an average accuracy of 83.3%. Social features such as number of conversations, number of lists and following to follower ratio, were utilised in the search to ensure a high level of accuracy in the sentiment classification task.

In (Hoang, et al., 2013) is presented a political-oriented tweets classifier and a Twitter user’s political affiliation classifier. Those methods in combination with a sentiment analysis tool were utilised to study the influence of political affiliation and message’s polarity on information sharing.

For the dataset collection, a set of seed users were selected; the set consisted of the major American politicians, political bloggers, mass media sites and some neutral users. Later, the set was expanded with users that were following three seed users, at least. Then, users’ tweets corresponding to August – October 2012 (presidential election period) were collected. Finally, a list of hashtags and political keyphases was defined to obtain a dataset containing only political oriented tweets. For the sentiment analysis task, it was employed the Stanford’s sentiment scoring API that implements a machine learning algorithm. The overall accuracy obtained for the political tweets was 77.9%.

The authors in (Cambria & Hussain, 2012) introduce an open-domain opinion mining and sentiment analysis engine based on a concept-based approach called sentic computing. This approach analyses text semantically to infer concepts and emotions associated with natural language opinions.

The proposed engine comprises the following four modules:
- Pre-processing module that is in charge of handling negation and preparing text for analysis.
- Semantic parser that converts text into multiple word concepts using an AffectNet and Isanette based lexicon. In addition, the parser provides a positive or negative connotation for each concept.
- Issanette module that detects semantics by exploiting the common sense knowledge base.
- AffectiveSpace module that infers sentsics by exploiting AffectNet and the Hourglass emotion model.

The dataset utilised for evaluation consisted of 2000 patient opinions associated to a category and the average F-measure achieved by the engine was 79.13%.

In (Bahrainian & Dengel, 2013) is proposed a hybrid sentiment analysis system that comprises three modules: a pre-processing module, sentiment feature generator module and a machine learning module for classification.

The pre-processing module includes the execution of steps such as URL and hashtag removal, slang replacement and division of text in smaller pieces according to punctuation characters. On the other hand, the sentiment feature generator is in charge of tagging all words with a score taking into account negation, intensifiers and diminishers. The SentiStrength lexicon is utilised for this task. Thus, the module generates a set of sentiment features such as number of positive/negative words, number of negation words followed by a positive/negative word, etc.

Finally, the machine learning module receives the set of features generated in the previous module and utilises a linear SVM algorithm to classify the tweets to classes. The dataset utilised for evaluation consisted of 940 tweets that were classified with an accuracy of 89.13%.

2.3 Summary

Various factors have made imperative the analysis of texts in order to extract people’s opinions. Social media, vast amount of data, new organisations’ requirements, etc, are examples of such factors.

On the other hand this task, often called sentiment analysis or opinion mining, has come a long way to be able to produce good results; from analysing entire documents to identifying specific opinionated portions, or from comparing single words to implementing complex natural language solutions.

However, this path has presented a lot of challenges that in part have contributed to such remarkable development and in part have limited its outcomes. Therefore, the research on this field is still necessary but promising at the same time. In this context, politics and government intelligence are enriching research areas due to people’s interest in expressing their opinion on such topics. It is so many related research works have been presented in application scenarios such as election result prediction and pattern identification in political events.
3 Project Initiation

3.1 Legal, Social, Ethical and Professional issues

Below are discussed significant implications that must be taking into consideration for proceeding appropriately at every stage of the project.

3.1.1 Legal Issues

It is clearly that the meaning of personal data has changed over time from including a full name to including an e-mail address or even a static IP address. In general any information that makes an individual identifiable can be considered personal data. In addition, any information that could be used to cause harm or discomfiture to an individual is considered as sensitive data (Minelli, et al., 2013).

Therefore, regarding this project, it is important to consider the types of data that will be managed. The tweets that will be utilised are posts that users themselves want to share, so it is publicly available. However, as it was mentioned above, some data can be considered as personal or sensitive information.

Thus, in order to comply with the principles dictated by the Data Protection Act and Global Privacy principles, only necessary data will be extracted and stored from Twitter. Additionally, any possible piece of data that is unique to an individual will be removed.

Although public Twitter accounts can be accessed anyway, the purpose of this project is not to identify or analyze any individual or citizen, so there is no necessity of storing data that identifies them.

3.1.2 Social Issues

At the time of dealing with social media data the social implications are considerable.

For instance, what if posts are misinterpreted, what if users are not willing to give its consent to be part of a research or what if the research process affects them in any way? All of these are some of the questions to reflect on (boyd & Crawford, 2012).

Regarding the given or not given consent for a research, it is interesting to notice the two recent revealed cases of experiments conducted by two social networking websites that have caused discomfort to users and generated criticism from various sectors.

In the first case, the article “Experimental evidence of massive-scale emotional contagion through social networks” presented by Adam D. I.Kramer, Jamie E. Guillory, and Jeffrey T. Hancock in June 2014, revealed that Facebook conducted a psychological experiment on its users during January 2012 (RT, 2014). Although Facebook users were not directly notified about this experiment, the authors indicated that all users agree to Facebook’s Data Use Policy and Terms of service, which state that their data may be used for data analysis and research.
The second case, involving the OkCupid dating and social network website, was revealed by Christian Rudder, its own co-founder (McBride, 2014). In July 2014, Rudder posted on a blog the description of the experiments they have used on its users. Again, OkCupid’s privacy policy indicates that customer’s information may be used for research and analysis.

Therefore, it is important to indicate that the proposed research does not try to experiment on people, but just perform a sentiment classification on what they have publicly shared. In fact, the Twitter dynamic of hearing and sharing opinions, as well as sharing what someone else has said by “retweeting” a post or marking as favourite, makes people aware that their posts will be seen and evaluated by everyone.

Regarding the misinterpretation of opinions, the research that will be carried out by this project actually tries to face this situation and evaluate the effectiveness of current techniques, so better tools can be provided.

Regarding possible ramifications or consequences, it is not erroneous to think that political interests may be affected, both from government’s side as well as opposition supporters. Therefore, in order to avoid actions that affect users, visualisations will be presented in a way that does not allow the identification of any user, but just allows distinguishing the proposed results.

3.1.3 Professional Issues

In order to avoid any professional issues, all processes will follow professional standards and guidelines. For instance, usability and accessibility guidelines will be considered at the time of developing the Web frontend. Additionally, as a professional it is important to understand and clearly state the limitations that may affect the project.

Therefore, one aspect that must be taken into consideration is the limitation of the data. Despite the number of tweets that may be utilised it is important to consider that Twitter users do not represent the entire population and that Twitter may not necessarily provide all public tweets but just a part of them (boyd & Crawford, 2012).

Other aspect that must be considered is the limitations of the analysis techniques. Sentiment analysis is one technique that cannot integrate the context of what is being said, consequently itself provides more of pattern recognition that anything else (Hinton & Hjorth, 2013).

Therefore, regarding this project, those limitations will be addressed during the research stage in order to find methods for dealing with them.

3.1.4 Ethical Issues

Some ethical implications that may arise have to be with manipulation of data. As a researcher, it is critical that results are not modified intentionally, not only because of the inappropriate conduct that this represents, but also because of the consequences that it may lead. Having this in mind, the data extracted as well as the data obtained after being analysed will be stored and presented as it is.
3.2 Research Method

3.2.1 Research Framework

The research intends to explore the current Sentiment analysis methods, tools and technologies; its application in a governmental context as well as the development of an integrated tool that allows visualising the results properly.

3.2.2 Motivation

Sentiment analysis is a complex and challenging subject that currently is the focus of attention of many researchers. This is due to its applicability in any area. Furthermore, the highly growing interest in social media data analysis has increased the importance of this area. Therefore, the relevance of the study of this topic is the main motivation of this research.

Additionally, considering that the social media environment has become a platform where people can communicate their opinions about political topics, governmental resolutions or election candidates, it represents a great opportunity to develop useful and interesting tools.

3.2.3 Research questions

- Sentiment analysis
  - What is the work that has been done so far?
  - What are the Sentiment analysis methods that can be utilised? And what are their performances?
  - What are the Sentiment analysis tools that are utilised?
  - What are the Sentiment analysis challenges to face for the proposed context?

- Data
  - What is the appropriate data storage for the tweets?
  - What is the process required to collect the tweets?

- Data Visualisation application
  - What are the options for visualising the analysis results?
  - What are the features that can be presented?
  - Which programming language, library or tool is more appropriate for the development?

3.2.4 Research plan

Below are presented the steps required to conduct the research during the proposed project, and that show the different stages of the research process.
<table>
<thead>
<tr>
<th>Engineering Method/Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  Define Research questions</td>
</tr>
<tr>
<td>2  Study existing solutions</td>
</tr>
<tr>
<td>2.1 Gather information</td>
</tr>
<tr>
<td>2.1.1 Literature survey</td>
</tr>
<tr>
<td>2.1.2 Search for existing solutions/products/tools</td>
</tr>
<tr>
<td>2.2 Elaborate literature review</td>
</tr>
<tr>
<td>3  Propose better/improved solution</td>
</tr>
<tr>
<td>3.1 Define assumptions</td>
</tr>
<tr>
<td>3.2 Define requirements/features</td>
</tr>
<tr>
<td>3.3 Define SW/tools to use</td>
</tr>
<tr>
<td>3.4 Define development methodology</td>
</tr>
<tr>
<td>4  Develop solution</td>
</tr>
<tr>
<td>4.1 Apply a development methodology</td>
</tr>
<tr>
<td>5  Analyse and Evaluate solution</td>
</tr>
</tbody>
</table>

Table 1: Research plan

3.3 Government’s Decisions

3.3.1 Ecuadorian Government Background

Current Ecuadorian Government started its administration in 2006. Since then, the incumbent president has been re-elected two times (2009 and 2013) for four-year terms. Considering that for the past ten years before this administration, the Ecuadorian political situation has been quite unstable due to forced changes in government, such unexpected stability has been analysed by several authors.

(Eichorst & Polga-Hecimovich, 2014) consider that such citizen support is due to various factors, including the increased social spending and a growing economy. However, they also point out that the government has faced some conflicts as well, mentioning a confrontation with national police in 2010 and accusations of restricting freedom of expression, as examples.

Meanwhile, (Becker, 2013) considers that the government stability is due in part to its policies and in part to people’s exhaustion after previous political crisis. Thus, the most significant decisions and policies implemented by the government are described and discussed extensively, highlighting the notorious benefits as well as the arisen criticisms.

One more noteworthy aspect that is pointed out is the creative use of media to keep the public informed (Becker, 2013). The incumbent administration has the habit of reporting on its activities to citizens in a weekly basis, via radio, television and live webcast. Government offices also share contents and information through diverse social media services as Facebook, Twitter, YouTube, Flickr and Ivoox.
In addition, some of these offices interact with citizens through Twitter by responding to requests or complains made by users.

3.3.2 Selection of Government’s Decisions

Below is described the process followed in order to select the three Ecuadorian Government’s Decisions to be analysed.

First, a list containing important government’s decisions or actions was elaborated. This list includes decisions widely promoted by the government, widely covered by the local and even international media, and widely evaluated by political analysts.

The list is presented below:
- Teacher evaluation
- Yasuni oil reserves exploitation
- Rejection of free-trade agreement with the United States
- U.S military forces from Manta’s air base
- Public Service Organic Law
- 2011 Referendum
- Mining concessions
- Education reforms
- Rift with Colombia (2008)
- Julian Assange asylum

Second, the following criteria were defined in order to select the three required decisions.
- Facility to determine a search criterion in Twitter.
- Transcendence
- Media coverage
- Amount of tweets generated

Then, taking into consideration the media coverage, there are two decisions that were covered internationally: the Julian Assange asylum and the Yasuni oil reserves exploitation. In addition, these two decisions may be referred easily as they contain names which facilitate the election of a search criterion in Twitter.

The Julian Assange asylum decision was certainly internationally transcendent as several countries were involved, while the Yasuni decision was more transcendent nationally. Lastly, an online Twitter search was carried out for each decision to confirm that there are a considerable amount of related tweets for analysis.

On the other hand, the rest of the decisions mentioned were mostly covered by local media and are transcendent in a national context. However, they can be referred using more than two words or in several different ways. Therefore, the 2011 referendum was considered as a good choice as it can be clearly referred with only two words for a Twitter search.
Likewise, an online Twitter search was carried out to verify the amount of related tweets available. For all the searches carried out, words in Spanish were used and a convenient range of dates were established to target the right results.

Below are presented the three decisions that were finally selected, including a brief description and the suggested range of dates for the Twitter search.

<table>
<thead>
<tr>
<th>Decision</th>
<th>Description</th>
<th>Reference dates</th>
<th>Search words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Julian Assange Asylum</td>
<td>WikiLeaks founder, Julian Assange, requested political asylum to the Ecuadorian Government entering the Ecuadorian embassy in London</td>
<td>19/06/2012: Asylum was requested</td>
<td>assange, asilo</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16/08/2012: Asylum was granted</td>
<td></td>
</tr>
<tr>
<td>Yasuni Oil reserves exploitation</td>
<td>The Ecuadorian government announces that the initiative of refraining from exploiting the oil reserves, located within a National Park called Yasuni, was abandoned. Therefore, the government decided to proceed with the responsible exploitation of the thousandth part of the Yasuni</td>
<td>15/08/2013: Decision announcement</td>
<td>yasuni</td>
</tr>
<tr>
<td>2011 Referendum</td>
<td>In 2011, the Ecuadorian Government called for a referendum that consisted of ten questions on different topics as constitutional, judicial and social issues</td>
<td>26/03/2011: Campaign start</td>
<td>consulta popular</td>
</tr>
<tr>
<td></td>
<td></td>
<td>07/05/2011: Referendum took place</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Ecuadorian Government's Decisions

3.4 Selection of Methods and Tools

3.4.1 Sentiment Analysis Method

As it was said previously, resources such as lexicons and labelled corpus are mainly available in English language. Given that languages may have different structural and grammatical features, sentiment analysis tools and applications tend to encounter many portability and suitability issues (Bosco, et al., 2013). These considerations will be taken into account as it is intended to analyse tweets in Spanish language.

There are social media search tools that offer sentiment features such as Social Mention\(^1\) and TwitterSentiment\(^2\). Other tools also offer off-the-shelf sentiment analysis services such as Blogmeter\(^3\) and some of them are freely available such as DatumBox\(^4\), Splunk Sentiment Analysis App and SentiStrength\(^5\) -that is free for academic purposes.

Although any of the options mentioned above could be used for the required sentiment analysis task, a refined approach will be selected. However, considering that the SentiStrength opinion mining program offers a Spanish version, this application will be utilised for comparison purposes.

---

\(^{1}\) http://socialmention.com
\(^{2}\) www.sentiment140.com
\(^{3}\) http://www.blogmeter.eu/
\(^{4}\) http://www.datumbox.com/
\(^{5}\) http://sentistrength.wlv.ac.uk/
Evaluating the results obtained by some related research works, it is possible to observe that the hybrid approaches have achieved better performance. Both studied approaches include pre-processing steps, but the approach that uses a supervised learning algorithm such as SVM outperforms the other method. Therefore, the hybrid approach presented in (Bahrainian & Dengel, 2013) will be utilised as a starting point for executing the sentiment analysis task. Thus, the mentioned approach will be improved in order to analyse the tweets in Spanish and will consider other features that may help to obtain a better accuracy.

On the other hand, the sentic computing approach proposed by (Cambria & Hussain, 2012) has convincing arguments to be taken into account. Both, the use of common sense knowledge bases and the implementation of a complex emotion model make this approach an appealing alternative. However, due to its complexity such method would be hardly implemented and no contributions would be presented. Hence, some tools introduced by this work will be considered in order to improve the chosen hybrid approach, which precisely seeks to combine proven solutions.

### 3.4.2 Tweets extraction

For the tweets extraction the Search API provided by Twitter will be used.

In order to send requests to the Twitter’s API and send the tweets to a convenient storage, it is possible to use out-of-the-box tools that wrap the Twitter’s API or create a personalized light application using a specific programming language.

In this case the second alternative has been chosen. Now, some of the programming language alternatives that have been considered are: Python, Ruby and Java.

As regards Python, (Russell, 2014) presents a good collection of examples for mining Twitter data with Python, which offers a specialised package for Twitter. On the other hand, several online resources present Ruby as a simple and reliable programming language that also has a specialised library and online documentation available for Twitter mining purposes. Finally, online Java examples show that it is possible to interact with the Twitter API by using Twitter4j package.

Considering the final environment where the solution could be implemented and the current programming language requirements in Ecuadorian public organisations, Java is the alternative that has been selected. However, the implementation of a Ruby application will not be discarded in case a performance improvement can be achieved.

### 3.4.3 Storage

Regarding the storage options, there are two general alternatives, to use a relational database or a NoSQL database. Considering the type of data that will be managed, both alternatives will be utilised.

First, the tweets will be stored in a NoSQL database that supports JSON-style documents.
Second, a relational database will be utilised to store the results of the sentiment analysis as well as any other resulting statistics.

Considering the Ecuadorian public organisation requirements, it is imperative that the tools to be utilised are free software.

In the case of the relational database, MySQL has been selected for its extended use among organisations.

As regards NoSQL storage, there are several alternatives; many of which were utilised for courses’ assignments. MongoDB, Cassandra, CouchDB, Hive and Pig are some of the examples.

More than considering the features that one or another may have, the selection was made considering the popularity and future demand of the product. Therefore, MongoDB has been selected for being the leading oriented-document NoSQL database according to Google searches and job postings (MongoDB, 2014).

3.4.4 Data Mining Tools and Visualisation

Considering time constraints, R has been selected for performing the necessary sentiment analysis steps, given that this tool was utilised for previous assignments with satisfactory results.

For visualisations, PHP and JSF have been considered as alternatives for developing the web application. However, considering the excellent results obtained in previous assignments, JSF has been chosen.

3.5 Development Methodology

Having several software development methodologies one must first understand that one methodology does not suits all type of projects (Ahmed, 2012). Then it must be considered that the size, development speed and information available at the beginning of a project are vital factors that conduct the appropriate selection of a development process.

In this case, the proposed project is small in terms of functionality as the main goals are to perform sentiment analysis and visualise the results through a Web interface. On the other hand, the complexity of the sentiment analysis task is high due to the limitations and challenges previously discussed.

Regarding the development speed, the project must deliver results as soon as possible given the current time constraints. And, as to the information availability at the beginning, it must be taken into consideration that the requirements were not set by real customers or users but rather elaborated based in real situations.

Therefore, an agile project delivery methodology is considered as appropriate, given the project features. In this context, the Dynamic Systems Development method (DSDM)
Atern is a methodology that provides a framework for delivering solutions at high speed and high quality (Tudor & Tudor, 2010).

### 3.5.1 Justification

Given that the aim of DSDM Atern is to deliver useful pieces as soon as possible but without losing quality, such development framework is considered an appropriate option among the agile methods.

Another determinant characteristic is that the lifecycle followed by this framework allows learning along the way as new requirements can be introduced or refined during the iterative process, and further work can be done at later stages.

Finally, considering the type of project that was proposed, its complexity and the target environment, the DSDM Atern is a favourable development methodology as the techniques and principles followed by the framework contribute to deliver fast solutions that can be enhanced and become a starting point for larger projects.

### 3.5.2 Techniques

Below is described how the DSDM techniques will be implemented:

- **MoSCoW Prioritisation:** This technique will be utilised to obtain a list of prioritised requirements starting from a general list derived from the project’s objectives. The fundamental and mandatory requirements will be satisfied by the proposed solution while the rest of the requirements will be left out for future work in case there is no more time.

- **Modelling:** In this case, the necessary models will be elaborated at the different phases of the DSDM Atern lifecycle. Especially, models derived from User-Centred Analysis and Design Techniques will be elaborated at Exploration and Engineering phases. In addition, UML and Database diagrams will also be incorporated when needed.

- **Iterative Development:** In this case, a vertical approach will be utilised for the iterations which will allow the solution evolution from the prioritised requirements to working products.

### 3.5.3 Lifecycle Model

Below is described how the DSDM Atern Lifecycle will be utilised to conduct an iterative, incremental and agile development.

- **Feasibility:** This phase was covered during the elaboration of the project’s proposal where a general view of the proposed solution was delivered, including the project’s objectives, plan and risk management.

- **Foundations:** At this phase, the prioritised list of requirements will be obtained, as well as high-level solution models and an Iteration Plan.
- Exploration: At this phase, an analysis of the solution’s requirements will be performed by utilising User-Centred Analysis techniques.
- Engineering: At this phase, User-Centred Design techniques will be utilised to deliver the necessary technical models and details that will guide the implementation task.
- Deployment: At this phase, the working products will be delivered and evaluated.

3.6 Summary

This section has presented the initial considerations and definitions that have been made for the proposed project.

First, it is presented the legal, social, ethical and professional issues that were considered for this project. Some of these considerations include the importance of protecting the identity of Twitter users, whose tweets will be analysed; as well as storing and presenting only the necessary data.

Other considerations remark the importance of managing the data responsibly and avoiding the alteration of any piece of data in order to manipulate results. Likewise, it is important to implement solution following standards so a high-quality product can be obtained.

Then, it is presented that the main reasons that motivates this research are the current relevance of Sentiment Analysis and the richness of the area for developing interesting tools. In this context, a research plan was presented which follows the Engineering Research Method.

This section also presents several selection processes that were conducted. First, the selection of Government’s decisions led to choose three relevant events for the country: The Assange’s asylum granting, the Yasuni oil exploitation approval and the 2011 referendum. Second, the selected Sentiment Analysis method corresponds to a hybrid approach that outperformed other single methods. And the selected development tools that correspond to free software alternatives such as: MongoDB, R and Java libraries.

In the case of the development methodology, the DSDM Atern framework was considered appropriate and an explanation of its use is given. In this context, it was presented a description of the techniques that will be utilised, as well as an explanation on how the DSDM lifecycle will be applied for this project.
4  Analysis and Design

4.1  Analysis Task

4.1.1  Assumptions Definition

-  Sentiment Analysis assumptions

  o  If the selected hybrid approach is applied as presented in (Bahrainian & Dengel, 2013), a lower accuracy will be obtained; this is because the domain and language of the tested dataset are different from the proposed ones.
   
  o  If the lexicon used for obtaining the sentiment features is improved considering the new language and domain, a similar or even higher accuracy will be obtained.
   
  o  If the lexicon used for obtaining the sentiment features is replaced for lexicon generated considering an emotion model, a similar or even higher accuracy will be obtained.
   
  o  The addition of more features to the sentiment features set may improve accuracy.
   
  o  If an extra classifier is added to the process pipeline, more tweets could be classified correctly. This classifier could be a Machine Learning algorithm.

-  Result set assumptions

  o  The polarities obtained will match with real results.

  o  If peaks are found evaluating the polarity values, these may indicate particular events.

  o  If the number of retweets and favourites for each tweet are taken into account, the trends could be more evident.

  o  Polarities shown along a timeline will allow identifying trends.

4.1.2  Requirements List

Below are listed the requirements of the proposed system.

1.  To obtain and store Twitter data corresponding to Government’s decisions
   
   a.  Extract data utilising pre-defined criteria

   b.  Store only the necessary data

   c.  Discard any piece of data that allows identifying users

2.  To perform Sentiment Analysis on each Government’s decision dataset
   
   a.  Perform analysis with an out-of-the-box tool for comparison purposes

   b.  Perform analysis with selected hybrid approach

   c.  Store analysis’ results for later presentation

3.  To evaluate methods’ performances
   
   a.  Classify manually a sample dataset

   b.  Calculate accuracy of each method utilised

   c.  Store evaluation results
4. To modify hybrid approach in order to obtain a higher accuracy
   a. Suggest improvements for hybrid approach
   b. Perform analysis with an improved hybrid approach
   c. Evaluate improved method performance

5. To present all results via a Web application
   a. Present polarity detection results for each decision and utilised method
   b. Present evaluation results for each decision and utilised method

4.1.3 Requirements Prioritisation
After executing the MoSCoW Prioritisation technique during the Foundations phase of the DSDM Atern lifecycle, the following prioritised requirements list was obtained:

<table>
<thead>
<tr>
<th>Prioritised Requirements List</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fundamental requirements</td>
<td></td>
</tr>
<tr>
<td>1 - Extract Twitter data utilising pre-defined criteria for each Government’s decision</td>
<td></td>
</tr>
<tr>
<td>2 - Store only the necessary data</td>
<td></td>
</tr>
<tr>
<td>3 - Discard any piece of data that allows identifying users</td>
<td></td>
</tr>
<tr>
<td>2 - Classify manually a sample dataset</td>
<td></td>
</tr>
<tr>
<td>3 - Perform sentiment analysis utilising an out-of-the-box tool</td>
<td></td>
</tr>
<tr>
<td>4 - Store out-of-the-box tool analysis’ results</td>
<td></td>
</tr>
<tr>
<td>5 - Calculate accuracy of out-of-the-box tool</td>
<td></td>
</tr>
<tr>
<td>3 - Present polarity detection results for each decision when utilising an out-of-the-box tool</td>
<td></td>
</tr>
<tr>
<td>4 - Present evaluation results for each decision when utilising an out-of-the-box tool</td>
<td></td>
</tr>
<tr>
<td>4 - Perform sentiment analysis with selected hybrid approach</td>
<td></td>
</tr>
<tr>
<td>5 - Store hybrid method analysis’ results</td>
<td></td>
</tr>
<tr>
<td>6 - Calculate accuracy of hybrid method</td>
<td></td>
</tr>
<tr>
<td>5 - Present polarity detection results for each decision when utilising selected hybrid method</td>
<td></td>
</tr>
<tr>
<td>6 - Present evaluation results for each decision when utilising selected hybrid method</td>
<td></td>
</tr>
<tr>
<td>6 - Suggest improvements for hybrid approach</td>
<td></td>
</tr>
<tr>
<td>7 - Perform analysis with a improved hybrid approach</td>
<td></td>
</tr>
<tr>
<td>7 - Store improved hybrid method analysis’ results</td>
<td></td>
</tr>
<tr>
<td>7 - Calculate accuracy of improved hybrid method</td>
<td></td>
</tr>
<tr>
<td>7 - Present polarity detection results for each decision when utilising selected hybrid method</td>
<td></td>
</tr>
<tr>
<td>7 - Present evaluation results for each decision when utilising selected hybrid method</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Prioritised Requirements
### 4.1.4 Iteration Plan

Based on the prioritised requirements list presented above, it is possible to define the following iterations in order to deliver useful and working parts for the proposed solution.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Prioritised Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fundamental requirements</strong></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Extract and store Twitter data</td>
</tr>
<tr>
<td></td>
<td>Manual classification of a sample dataset</td>
</tr>
<tr>
<td></td>
<td>Sentiment Analysis and evaluation using an out-of-the-box tool</td>
</tr>
<tr>
<td></td>
<td>Results visualisation via Web application</td>
</tr>
<tr>
<td><strong>Mandatory requirements</strong></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Sentiment Analysis and evaluation using selected hybrid method</td>
</tr>
<tr>
<td></td>
<td>Results visualisation via Web application</td>
</tr>
<tr>
<td><strong>Suggested requirements</strong></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Hybrid approach improvement</td>
</tr>
<tr>
<td></td>
<td>Sentiment Analysis and evaluation using improved hybrid method</td>
</tr>
<tr>
<td></td>
<td>Results visualisation via Web application</td>
</tr>
</tbody>
</table>

As it is shown above, the iterations will be performed during the Engineering phase in order to deliver a complete solution incrementally.

### 4.1.5 Proposed solution

Below is presented a general perspective of the proposed solution given the general requirement list.

![Solution general perspective](image-url)
In addition, a solution’s perspective derived from the prioritised requirements list is presented below:

![Diagram of iterative process]

**Figure 2: Solution General View in the iterative process**

### 4.2 Web Application Analysis

#### 4.2.1 Web application overview

The proposed Web application will allow presenting the results of the Sentiment Analysis task for each Government’s decision, as well as the evaluation results corresponding to each method utilised for such analysis.

Therefore, the user will be able navigate through each decision and learn the resulting Twitter sentiment. In addition, for didactic purposes, the user will be able to inspect the results obtained by the different methods utilised in the project. In this case, standard charts will be utilised.

On the other hand, the accuracies and other evaluation parameters will be presented as well for all the methods and for each decision, so the user can assess the validity of the results obtained in the sentiment analysis task. In this case, summary tables will be utilised to present these values.

#### 4.2.2 User / Audience analysis

##### 4.2.2.1 Potential Users

The following groups of users were identified as potential users for the proposed Web application.

- Data scientists
- Female and male Ecuadorian Communication Secretary personnel
Female and male Government officers
Female and male communication professionals
Female and male political analysts
Female and male journalists
People interested in politics
Adult Twitter users

4.2.2.2 Roles

Below is listed a classification of users according to the tasks that can perform through the application:
- Common user: This type describes a user who is only interested in exploring the sentiment polarities for a particular subject, in this case a Government’s decision.
- Data specialist user: This type describes a user who is interested in examining the performance of different Sentiment Analysis methods.

4.2.2.3 Characteristics

The characteristics listed below were identified for the potential users.
- Users have enough computing skills for navigating web pages.
- Users have at least a basic understanding on graphical and tabular representations of data.
- Users probably participate somehow on social media.
- Majority of users will use the application during working time.
- Users will access to the application through web browsers such as Google Chrome, Mozilla Firefox, Internet Explorer and Safari.

4.2.2.4 Differences

- Data scientists differ from other groups in the type of information they require that is related to the performance of the Sentiment analysis module.
- Ecuadorian Communication Secretary Personnel and Government officers differ from other groups in the purpose of use. They are interested in people’s opinion on specific subjects so they can elaborate strategic actions.
- Communication professionals and journalists differ from other groups in the purpose of use. They are interested in people’s opinion for informing the public.
- People interested in politics and adult Twitter users differ from other groups in the purpose of use. They are interested in people’s opinion for personal acquaintance.

4.2.2.5 Persona representations

In order to bear in mind the characteristics of the users who will potentially utilise the Web application, three “Persona” representations were created.
PERSONA: SILVIA

Silvia Lopez
“What is behind the data?”

Age: 28
Data scientist
Working on a Big Data application

Silvia is a 28 year-old female who works for a Business Intelligence Software company. She likes practicing sports during the weekend. She works as an associated consultant so she does not have a defined working time, therefore she sometimes has its own development projects. She owns a tablet but prefers working on her laptop. She is specialist in data integration and business analysis. She is also an application developer. She has skills on data analysis, statistics and probabilistic analysis skills. She is always interested in using new tools for developing business intelligence solutions.

Figure 3: Persona Silvia

PERSONA: FERNANDO

Fernando Alvarado
“Responsible media is what society needs”

Age: 35
Communication Secretary functionary
Working in a communication campaign

Fernando is 35 years old. He works for the Ecuadorian Communication Secretary at the planning area. He is a communication professional. He also assists the organisation of press conferences and discussion groups for Government authorities. He is very active in social media websites and supervises the official communications published by the secretary. He manages his own personal social media accounts. He likes to listen to music while he is working, as well. Sometimes, he works on his reports late at night. He has intermediate skills on computing and Internet.

Figure 4: Persona Fernando
4.2.3 Workflow analysis

The following diagram shows the main interactions that the web application will have.

**PERSONA: EVA**

Eva Garcia
“Revolutionary”

Age: 41
Journalist
Working on a book.

Eva is 41 years old. She works for a national newspaper.
She writes different kind of articles but mostly political articles. She sometimes travels to make special interview for an exclusive section.
All days she supervises the leading articles that will be published.
Sometimes, she participates as a political analyst in local television news.
She has been very critical about certain Government’s policies.
She wears glasses and complains about having headaches after working several hours
She is a intermediate Internet user and is very active in Twitter.

**Figure 5: Persona Eva**

**Figure 6: Main interactions**
4.3 Design

4.3.1 Web application pages

This section presents mock-ups for the pages of the proposed Web application. These mock-ups were created using the open-source desktop application, Pencil Project\(^6\).

4.3.1.1 Home page

The Home page presents a main menu with all the links to the existent pages. Taking into consideration the physical limitations of Persona 3, it will be necessary to ensure that the contrast between the background and text colours is high.

![Home page](image)

**Figure 7: Home page**

4.3.1.2 Opinion page

When a user selects the Opinion option of a decision, the Opinion page is presented. This page contains the results of the Sentiment Analysis for the decision selected. These results are presented using appropriate charts. In addition, the page includes a left-side menu that allows the navigation through other decisions.

It is also possible to navigate to the respective Methods Assessment page from this point by selecting the corresponding tab header.

---

\(^6\) Pencil Project Web Site: http://pencil.evolus.vn/
4.3.1.3 Methods assessment page

When a user selects the Methods assessment option of a decision, the Methods assessment page is presented. This page contains the results of the evaluation of the methods performance that were utilised for the Sentiment Analysis. These results are presented using appropriate charts. In addition, the page also includes the menu that allows the navigation through other decisions.

It is also possible to navigate to the respective Opinion page from this point by selecting the corresponding tab header.

Considering the computing skills of the potential users described by the Persona representations, it was noticed that the use of tabs for showing the results will favour the page disposition and navigability.
4.3.2 Database

The following diagram presents how the collections will be organised in the document-oriented database for the Twitter data. The essential document fields are also specified, although it is possible that some vary since the semi-structured feature will be exploited.

The following diagram presents the entity relational model for the database that stores the evaluation results.
In addition, a detailed diagram showing the keys and column names for the proposed relations is shown below:

![Diagram showing the keys and column names for the proposed relations](image)

**Figure 12: Relational Database tables**

### 4.3.3 Font for Web pages

The font size will be set using CSS. The body text will be left to the default browser size, while the headings and links will use “ems” units. This way, the text scalability is assured as recommended in (Lynch & Horton, 2011).
The selected font family corresponds to a web safe combination as indicated in (W3Schools, 2014). The first typeface is the preferred type and the rest corresponds to a popular typeface as well as generic families. The defined family is as follows:

{Georgia, Times New Roman, Times, Serif}

4.4 Summary

This section has presented the application of several Analysis and Design techniques that allowed obtaining the necessary details to initiate the implementation task. Such techniques were performed as part of the Foundations and Exploration phases of the DSDM Atern development framework.

First, the Task analysis allowed obtaining the high-level requirements specification and a general view of the proposed solution, all as part of the Foundations phase.

Then, the Web application analysis was presented as part of the Exploration phase. For this, user-centred analysis techniques were utilised to obtain the characteristics of the potential users for the required Web application. Later, these characteristics were taken into consideration during the web pages design; and mock-ups of the main web pages were created as well.

The database design models for the document-oriented and relational databases were also presented in this section.

Finally, the points taking into consideration for defining the font size y typefaces were also mentioned.
5 Implementation

5.1 First iteration

5.1.1 Development environment setup

For the development process and implementation of the proposed solution, the following local environment was prepared.

A virtual machine on Virtual Box was created, with a Memory Base of 5GB and Storage maximum capacity of 50GB. The virtual machine hosts an OpenSUSE 13.1 operating system with a GNOME desktop environment.

For data storage, MongoDB document-oriented database and MySQL relational database were installed.

For the programming and processing tasks, Netbeans IDE for developing Java and JSF applications, was installed. In addition, the statistical computing and graphics environment, R, was installed.

For the web application deployment, Glassfish application server was installed. While for the execution of some periodic tasks, the Linux Cron job scheduler was utilised to run the java applications.

![Diagram](image)

**Figure 13: Development environment**
5.1.2 Twitter data extraction and storage

First, a Java application was developed to obtain and store the tweets in MongoDB. For accessing the database, the Java Mongo Driver provided by MongoDB was utilised. Then, for consuming the Twitter Search API, the Twitter4j Java library was added to the application.

Previously, in order to obtain the required credentials for calling the service, a Twitter application was created for the selected Twitter account. This way, the application-only authorization was chosen as it does not require that the user’s login details are shared. The only details that should be specified in the Java application are the API keys and tokens provided by the Twitter application.

Thus, utilising the methods provided by the Twitter4j library, the required search criteria were specified in order to obtain the Twitter datasets.

However, due to Twitter API restrictions it is not possible to obtain tweets published over a week ago from current date. Therefore, the use of the Search API was discarded as the required data dates back to 2011. Even so, such application was implemented and tested.

In consequence, it was decided to utilise the Twitter Advanced Search functionality which allows specifying criteria, a date range and a language for performing a search. The criteria that were specified consisted of one or two keywords, a range of dates, Spanish language selection and inclusion of re-tweets.

Then the displayed tweets were saved into text files. However, in order to obtain the major quantity of tweets, it was necessary to extract the tweets specifying a day, rather than a date range.

Thus, the following table illustrates the quantity of tweets obtained for each decision and the date range to which they belong.

<table>
<thead>
<tr>
<th>Decision</th>
<th>Keyword</th>
<th>Date range</th>
<th>Number of tweets</th>
<th>Collection name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Julian Assange's Asylum</td>
<td>assange asilo</td>
<td>16/08/2012 - 01/09/2012</td>
<td>4765</td>
<td>assange</td>
</tr>
<tr>
<td>Yasuni Oil Exploitation</td>
<td>yasuni</td>
<td>15/08/2013 - 01/09/2013</td>
<td>6938</td>
<td>yasuni</td>
</tr>
<tr>
<td>2011 Referendum</td>
<td>consulta popular</td>
<td>25/03/2011 - 07/05/2011</td>
<td>4478</td>
<td>consulta</td>
</tr>
</tbody>
</table>

Table 5: Tweets extracted summary

Then, a java application called “tw_extraction” was created in order to read the text files and insert each tweet into the MongoDB database. For this purpose, the text files follow a basic structure: each tweet occupies two lines. The first line contained the creation date, which was extracted and formatted conveniently, and the second line contained the text of the tweet.

For inserting the tweets the application utilises the DBOBJECT interface provided by the Java Mongo Driver that allows saving custom objects. Basically, the data that was inserted were the tweets with its corresponding creation dates. Thus, taking advantage of Mongo features
such as automatic collection creation and id generation, three Twitter datasets were stored, one for each decision.

The followed process is illustrated by figure below:

5.1.3 Manual sentiment classification

In order to perform a manual classification of a part of the tweets according to the opinion expressed, a simple web application was developed and finally utilised.

The application requires the user selects a collection and enters a date. Thus, the application requests all tweets in the collection that correspond to the specified date and list them in a web page. The JSF component library Primefaces\(^7\) was utilised in this case.

Once listed, alongside each tweet, the option of assigning a polarity to each tweet was presented. This is shown below:

\(^7\) Primefaces website: http://primefaces.org/
5.1.4 Sentiment Analysis module

In this first iteration, the sentiment analysis module comprises the implementation of a pre-processing module and the implementation of an out-of-the-box sentiment analysis tool. This way, the tweets will be pre-processed as needed and later classified in an automated process.

As a first step, the pre-processing module is formed by the combination of a Java application called “tw_r_module” and a user-defined R script. In this case, the application obtains all tweets corresponding to a collection; then, the tweets are written to a text file that is accessible from the R environment. Later, the application utilises the JRI\(^8\) Java/R interface which allows running R functions inside Java applications for making a call to a user-defined R function that reads the tweets in the file and execute a pre-processing task. In this case, the required task was the removal of URLs from all tweets.

Ultimately, the Java application reads again the processed tweets provided by the R function and store them back into MongoDB database adding them as a new field under each original tweet. A screenshot showing the results of this process is presented below:

8 http://rforge.net/JRI/
In addition, the process and required interactions for the pre-processing task are illustrated by the following figure:

![Diagram of pre-processing module]

**Figure 18: Pre-processing module**

Then, as a second step, a sentiment analysis tool that is able to handle text in Spanish language was selected among the following options:

- **Bitex API**: Evaluation free and commercial tool that can be integrated by calling an API. This tool uses linguistic analysis based on grammars that allows identifying phrases and their dependencies. The evaluation free version limits the API calls to 30 days, 1000 calls per day, 8KB sent per call and maximum 2MB sent per month.
- **Apicultur**: Is an API platform that offers linguistic functions. The sentiment analysis API is exclusively a tool for analysis in Spanish. It is a commercial tool that offers free €20 monthly which allows performing a maximum of 100000 API calls. It only evaluates the first 140 characters sent.
- **Sentiment140**: Commercial version as well free version previous registration, that uses a Maximum Entropy Classifier and it handles request for analysis in Spanish languages as well.
- **Atribus**: Only commercial version.
- **Textalytics**: Free and commercial version. 500000 credits free which is equivalent to a maximum of 500000 words that can be analysed monthly.

Thus, taking into consideration the main language target, the number of tweets that can be analysed without cost and the implementation difficulties, the Apicultur API for sentiment analysis called “stmtlk 1.0.0”, was chosen.
Then, a Java application called “sa_box_client” was created in order to call the Apicultur API via a Web service request.

First, the application extracts 20 tweets not previously analysed from the MongoDB database. This is because the API restricts the number of tweets that can be sent per minute to 20. Then, for each tweet, the text is sent to the Web service by using the lightweight HTTP library for Java called “Unirest”, and whose responses are automatically parsed into JSON responses. The response sent by the API is comprised of four fields: processed text, polarity, intensity and certainty. The polarity element classifies each tweet as positive, negative or neutral. However, all the response content is inserted into the MongoDB, under the respective original tweet.

Then, in order to execute the application every minute, the Linux cron job scheduler was utilised. The following figure shows the log registered during the execution of the mentioned application.

![Figure 19: sa_box_client application log](image1)

In addition, the following figure illustrates the interactions described above:

![Figure 20: Use of Sentiment analysis tool](image2)
5.1.5 Data Visualisation module

The Web application “opinion_decision” was created for presenting the sentiment analysis results as well as the method assessment results. As first step, the tables were created in MySQL relational database having the following specifications:

<table>
<thead>
<tr>
<th>DECISION</th>
<th>PK</th>
<th>decision_id</th>
<th>INT</th>
</tr>
</thead>
<tbody>
<tr>
<td>decision_name</td>
<td></td>
<td>VARCHAR(30)</td>
<td></td>
</tr>
<tr>
<td>decision_description</td>
<td></td>
<td>VARCHAR(300)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>METHOD</th>
<th>PK</th>
<th>method_id</th>
<th>INT</th>
</tr>
</thead>
<tbody>
<tr>
<td>method_name</td>
<td></td>
<td>VARCHAR(30)</td>
<td></td>
</tr>
<tr>
<td>method_description</td>
<td></td>
<td>VARCHAR(200)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MEASURE</th>
<th>PK</th>
<th>measure_id</th>
<th>INT</th>
</tr>
</thead>
<tbody>
<tr>
<td>measure_name</td>
<td></td>
<td>VARCHAR(30)</td>
<td></td>
</tr>
<tr>
<td>measure_description</td>
<td></td>
<td>VARCHAR(200)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PERFORMANCE</th>
<th>PK, FK1</th>
<th>decision_id</th>
<th>INT</th>
</tr>
</thead>
<tbody>
<tr>
<td>method_id</td>
<td></td>
<td>INT</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PERFORMANCE</th>
<th>PK, FK2</th>
<th>method_id</th>
<th>INT</th>
</tr>
</thead>
<tbody>
<tr>
<td>measure_id</td>
<td></td>
<td>INT</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PERFORMANCE</th>
<th>PK, FK3</th>
<th>measure_id</th>
<th>INT</th>
</tr>
</thead>
<tbody>
<tr>
<td>value</td>
<td></td>
<td>DECIMAL(9,4)</td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Tables in relational database

Later, entity beans were generated for accessing the relational database from the Web application. Required session beans and JSF Managed Beans were also defined. Then, the main Web pages were built up. First, two CSS files were defined for controlling the page layout and styles.

For creating the charts, the Primefaces JSF component library was utilised. The types of charts selected were a stacked bar graph for the opinion results while a linear chart was selected for representing the number of tweets over time.

The following screenshot presents the main menu defined for the Home Page.
For this first iteration the sentiment obtained by the manual and the automated sentiment classifications are presented, for each decision. The following screenshots shows such visualisations for the 2011 Referendum decision.

![Opinion results for manual classification](image1)

**Figure 22: Opinion results for manual classification**

![Opinion results for classification using “Apicultur” API](image2)

**Figure 23: Opinion results for classification using “Apicultur” API**

A linear graph is also included for showing the number of tweets per day. This is shown below:
Finally, in order to present the method assessment results, a data table provided by “Primefaces” library was utilised. Below is presented the page showing such results for one of the decisions.

![Figure 25: Apicultur method assessment results](image)

5.1.6 Assessment module

The measures taking into consideration for assessing the method used by the out-of-the-box sentiment analysis classifier have been used by (Khan, et al., 2014) and are the following:
- Precision Class A: Proportion between correctly Class A classified cases (by the classifier) and all Class A classified cases (by the classifier).
- Recall Class A: Proportion between correctly Class A classified cases (by the classifier) and manual Class A classified cases.
- F-measure Class A: Harmonic mean of both precision and recall for Class A
- Method Accuracy: Proportion of all correctly classified results from all given data.

For expressing these measures mathematically, a confusion matrix is defined as follows (Khan, et al., 2014):

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Predicted Class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>tA</td>
</tr>
<tr>
<td>B</td>
<td>BA</td>
</tr>
<tr>
<td>C</td>
<td>CA</td>
</tr>
</tbody>
</table>

![Figure 26: Confusion matrix](image)

Where A, B, C are three different classes; “tA”, “tB” and “tC” are the corrected classified cases for such classes and the rest of elements are the incorrectly classified cases. For instance “BA” represents the number of “B” cases classified incorrectly as “A” cases.

Thus, the measures are expressed mathematically as follows:

\[
\text{Precision}_A = \frac{tA}{tA + BA + CA} \quad \text{Precision}_B = \frac{tB}{tA + tB + CB} \quad \text{Precision}_C = \frac{tC}{tA + BC + tC}
\]

\[
\text{Recall}_A = \frac{tA}{tA + AB + AC} \quad \text{Recall}_B = \frac{tB}{BA + tB + BC} \quad \text{Recall}_C = \frac{tC}{CA + CB + tC}
\]

\[
F\text{Measure(Class)} = \frac{2 * \text{Precision(Class)} * \text{Recall(Class)}}{\text{Precision(Class)} + \text{Recall(Class)}}
\]

\[
\text{Accuracy} = \frac{tA + tB + tC}{tA + AB + AC + BA + tB + BC + CA + CB + tC}
\]

Considering the Negative, Neutral and Positive polarities as classes, a confusion matrix was generated for them and the above formulas were utilised to calculate the mentioned measures.

An R function was defined for performing these calculations and the values obtained were stored in the respective performance record of the relational database. Later these values are retrieved and presented by the Web application as shown in the previous section.
5.2 Second iteration

5.2.1 Sentiment Analysis module

For the second iteration, the sentiment analysis module comprises the implementation of the selected hybrid method.

The method requires the implementation of three main modules: pre-processing, sentiment generator and machine learning classifier. The following diagram shows the interaction between the modules.

![Diagram of hybrid method implementation modules]

The pre-processing module generates outcome data that is saved into the MongoDB database to be later utilised by the Sentiment Feature module. In the same way, the pre-processing module generates outcome data that is saved into the MongoDB database to be later utilised by the Machine Learning Classifier.

5.2.1.1 Pre-processing module

The pre-processing module takes the available tweets of a collection for preparing them for the next steps. The pre-processing tasks that are executed comprise the following:

- URL Removal: the module looks for the presence of one or more URLs in each tweet, so they are removed.
- “#” Removal: the module looks for the # characters to be deleted.
- Usernames replacement: Any username mentions inside the tweets are replaced by the “ATUSER” expression.
- Slangs or abbreviations are replaced by their actual phrase equivalent. For this purpose, a slang/abbreviation dictionary was manually built by searching common Spanish terms and its equivalents, over the Internet.
- The sentiment target, in this case, each decision keyword was replaced by the “TARGET” expression.

For implementing these steps, a method called pre-process was added to the existent Java application “tw_r_module”. This method calls an R function which returns the processed the tweets. The results of executing this method are shown in the following figure:
5.2.1.2 Sentiment Feature Generator

The sentiment feature generator utilises the pre-processed tweets to generate sentiment features by computing different characteristics of each tweet. The following diagram shows the steps and outcomes during this process.

First, a polarity lexicon is used for determining the polarity of the words present in each tweet. The polarity lexicon consists of word stems. Each word in the tweet that starts with a stem in the lexicon, it is replaced by their respective polarity ([1] = positive or [-1] = negative). In addition, emoticons, negation words and booster words are also replaced by their respective polarity, for which a polarity lexicon of these expressions and words are also utilised.

The result of such tagging process is shown below:
Later the tagged tweets are used to calculate the sentiment features that are described below:

- Feature 1: Tweet sentiment score
- Feature 2: Number of positive words
- Feature 3: Number of negative words
- Feature 4: Number of negation words
- Feature 5: Number of negation words followed by a positive word
- Feature 6: Number of negation words followed by a negative word
- Feature 7: Number of positive words followed by the target
- Feature 8: Number of negative words followed by the target
- Feature 9: Number of negation words followed by the target
- Feature 10: Number of positive words followed by a positive word
- Feature 11: Number of negative words followed by a negative word
- Feature 12: Number of target words followed by a positive word
- Feature 13: Number of target words followed by a negative word

For performing these calculations, the Apache Commons Lang\(^9\) Java library was utilised to count the occurrences of such sequences. Thus, the following figure shows the results obtained after executing this process:

\(^9\)http://commons.apache.org/proper/commons-lang/
5.2.1.3 Machine Learning Classifier

Finally, the features values obtained for each tweet are utilised to train and test an SVM classifier. Thus, for training the classifier the manually classified tweets corresponding to the Assange’s decision dataset were utilised. The multi-class classification extension included in the “lib-svm” R library is required as three classes (positive, negative and neutral) are intended to be obtained. Once the model is obtained, the rest of the tweets were classified and the results obtained, were stored into the MongoDB database.

5.2.2 Data Visualisations module

For the second iteration, the data visualisation module is extended by presenting the sentiment analysis results for the Hybrid method as well as the corresponding method assessment results. The following figures present such visualisations:

Figure 31: Sentiment Features generation results
5.3 Third iteration

5.3.1 Sentiment Analysis module

For the third iteration it was proposed to introduce modifications in order to improve the hybrid method’s accuracy.

Thus, two steps were introduced to the sentiment feature generation process. First, a sentiment phrase polarity lexicon was obtained from the lexicon delivered by (Saralegi X., 2013). This was utilised for searching and replacing such phrases for its equivalent polarity.

This is done with the purpose of finding sentiment expressions formed by more than one word, which have its own polarity as a whole.
Second, after all polarity equivalents have been inserted in each tweet, the texts are processed again by removing Spanish stop-words and unnecessary whitespaces. The purpose of these extra steps is to improve the possibilities of finding tag combinations that were not possible for the presence of stop-words between two tags or extra whitespaces left by previous replacements.

The following figure presents the results of this process:

![Figure 34: Tag results after introduced modifications](image)

5.3.2 Data Visualisations module

For the third iteration, it has been added the sentiment analysis results visualisations corresponding to the Modified Hybrid method, as well as the method’s assessment results. This is shown below:

![Figure 35: Sentiment Analysis results using Modified Hybrid method](image)
5.4 Summary

The chapter has described the implementation of the selected sentiment analysis method and the development of the web application utilised to present the results.

Following the development methodology plan for the deployment phase, the working parts of the solution were implemented incrementally throughout three iterations.

The first iteration delivered the majority of the working product by establishing the development environment, the sentiment analysis infrastructure and the Web application. In this case, the manual sentiment analysis classification was implemented as well as the use of an online sentiment analysis API for Spanish.

Then, the second iteration focused on the implementation of the selected hybrid method. Such implementation required the definition of R scripts which were invoked by distinct Java Applications; these applications in turn were in charge of storing the results of each step in the process into the MongoDB database. Then, the Web application was modified to present the results obtained by the hybrid method polarity classification.

Finally, during the third iteration, the hybrid method analysis process was modified by the inclusion of new steps that could improve the accuracy obtained at first. Consequently, the Web application was modified to present the results obtained by the modified hybrid method.
6 Results and Evaluation

6.1 Results

After concluding the deployment phase, the following results were obtained.

6.1.1 Julian Assange Asylum decision

- Number of tweets per decision and per day: Out of a total of 4765 tweets collected over a period of 17 days, it is observable that the majority of tweets regarding the topic correspond to the first days after the decision announcement. In addition, it is also observable the occurrence of two peaks on 24 August 2012 and 30 August 2012. According to a chronology presented by (DiarioLibre.com, 2014) those dates correspond, respectively, to the OEA (Organisation of American States) resolution of supporting Ecuadorian Government’s decision and Wikileaks spokesman declaration, affirming the existence of a secret process against Assange.

![Figure 37: Assange decision - Number of tweets per day](image)

- Manual Classification: A total of 517 tweets were manually classified as neutral, positive and negative tweets. Almost all of these tweets correspond to the day when the asylum granting decision was announced. The negative and positive opinions slightly differ in quantity.

![Figure 38: Assange decision - Manual Classification](image)
• Apicultur sentiment analysis: The majority of tweets were classified as negative.

![Figure 39: Assange decision – Apicultur API analysis](image)

• Hybrid method analysis: Among the tweets not classified as neutral, a negative tendency can be noticed; especially the day after the decision was announced.

![Figure 40: Assange decision - Hybrid method analysis](image)

• Modified Hybrid method analysis: A similar distribution to the previous analysis for the negative side is presented; however, the positive side presents a more natural distribution.

![Figure 41: Assange decision - Modified Hybrid method analysis](image)
• Methods assessment results and comparison: Comparing the Hybrid method and the modified Hybrid method, there are slightly differences between them performance-wise, resulting that the modified method was outperformed by the initial hybrid method. On the other hand, comparing the Apicultur API and the Hybrid method implementation, it is noticeable that the hybrid method shows a better accuracy. This is due to its better performance in classifying neutral and positive tweets; while in the case of classifying negative tweets, the Apicultur API shows a better performance.

<table>
<thead>
<tr>
<th>Assessment Measures</th>
<th>Apicultur</th>
<th>Hybrid Method</th>
<th>Hybrid Method Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall Positive</td>
<td>3.9474%</td>
<td>5.2632%</td>
<td>6.5789%</td>
</tr>
<tr>
<td>Precision Positive</td>
<td>17.6471%</td>
<td>36.3636%</td>
<td>33.3333%</td>
</tr>
<tr>
<td>F-measure Positive</td>
<td>6.4517%</td>
<td>9.1955%</td>
<td>10.9889%</td>
</tr>
<tr>
<td>Recall Negative</td>
<td>93.0556%</td>
<td>8.3333%</td>
<td>8.3333%</td>
</tr>
<tr>
<td>Precision Negative</td>
<td>13.9293%</td>
<td>13.3333%</td>
<td>15.0000%</td>
</tr>
<tr>
<td>F-measure Negative</td>
<td>24.2314%</td>
<td>10.2564%</td>
<td>10.7143%</td>
</tr>
<tr>
<td>Recall Neutral</td>
<td>2.7624%</td>
<td>88.8889%</td>
<td>88.3469%</td>
</tr>
<tr>
<td>Precision Neutral</td>
<td>83.3333%</td>
<td>71.1497%</td>
<td>70.5628%</td>
</tr>
<tr>
<td>F-measure Neutral</td>
<td>5.3475%</td>
<td>79.0362%</td>
<td>78.4597%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>15.6863%</td>
<td>65.3772%</td>
<td>65.1838%</td>
</tr>
</tbody>
</table>

Figure 42: Assange decision - Methods assessment

6.1.2 Yasuni exploitation decision

• Number of tweets per decision and per day: Out of a total of 6938 tweets collected over a period of 18 days, it is observable that the quantity of tweets per day is uniform except for one day that the quantity decreases notably after a high peak. This peak coincides with an event carried out by decision opponents.

Figure 43: Yasuni decision - Number of tweets per day

• Manual Classification: A total of 376 tweets were manually classified by its polarity. All of these tweets correspond to the day when the decision was
announced. The quantity of negative tweets is noticeable higher than the positive ones and neutral ones.

Figure 44: Yasuni decision - Manual Classification

- Apicultur sentiment analysis: For each day the quantity of negative tweets exceeds the quantity of positive tweets.

Figure 45: Yasuni decision – Apicultur API analysis

- Hybrid method analysis: The tweets classified as negative exceed the positive tweets in almost all cases, and in some cases just for a slightly difference.

Figure 46: Yasuni decision - Hybrid method analysis

- Modified Hybrid method analysis: In just one case the quantity of positive tweets exceeds the negative ones clearly. In the rest of the cases, the quantity of negative tweets is higher.
Methods assessment results and comparison: Comparing the Hybrid method and the modified Hybrid method, there are almost no differences between their performances. On the other hand, comparing the Apicultur API and the Hybrid method implementation, the Apicultur API shows a better accuracy. This is due to its better performance in classifying negative and positive tweets; while in the case of classifying neutral tweets, the Hybrid methods show a better performance.

### Figure 48: Yasuni decision - Methods assessment

<table>
<thead>
<tr>
<th>Assessment Measures</th>
<th>Apicultur</th>
<th>Hybrid Method</th>
<th>Hybrid Method Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall Positive</td>
<td>24.2424%</td>
<td>10.5283%</td>
<td>2.6316%</td>
</tr>
<tr>
<td>Precision Positive</td>
<td>10.8108%</td>
<td>6.6667%</td>
<td>1.7241%</td>
</tr>
<tr>
<td>F-measure Positive</td>
<td>14.9533%</td>
<td>8.1633%</td>
<td>2.0833%</td>
</tr>
<tr>
<td>Recall Negative</td>
<td>49.5238%</td>
<td>16.4502%</td>
<td>18.6147%</td>
</tr>
<tr>
<td>Precision Negative</td>
<td>63.0303%</td>
<td>62.2951%</td>
<td>64.1791%</td>
</tr>
<tr>
<td>F-measure Negative</td>
<td>55.4667%</td>
<td>26.0274%</td>
<td>28.8590%</td>
</tr>
<tr>
<td>Recall Neutral</td>
<td>31.7073%</td>
<td>85.0467%</td>
<td>83.1776%</td>
</tr>
<tr>
<td>Precision Neutral</td>
<td>30.2326%</td>
<td>35.6883%</td>
<td>35.4582%</td>
</tr>
<tr>
<td>F-measure Neutral</td>
<td>30.9524%</td>
<td>50.2763%</td>
<td>49.7207%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>42.4615%</td>
<td>35.3723%</td>
<td>35.3723%</td>
</tr>
</tbody>
</table>

6.1.3 2011 Referendum

- Number of tweets per decision and per day: Out of a total of 4478 tweets collected over a period of 44 days, it is observable that the quantity of tweets per day increases as the date is closer to the referendum day.
Manual Classification: A total of 1252 tweets were manually classified by its polarity. These tweets correspond to different days of the selected period, especially to the first and last days. The quantity of negative tweets is higher than the quantity of positive ones; however the majority of tweets were classified as neutral.

Apicultur sentiment analysis: According to this analysis, for the majority of days, the quantity of positive tweets exceeds the quantity of negative tweets.

Hybrid method analysis: In this case, there is still a major quantity of tweets classified as positive, and there are some cases where there are no positives cases.
occurrences. It is also noticeable that the majority of tweets were classified as neutral.

- **Modified Hybrid method analysis:** The results are similar to the ones obtained by the original Hybrid method

- **Methods assessment results and comparison:** Comparing the Hybrid method and the modified Hybrid method, there are little differences between their performances. The original Hybrid method outperforms the modified one, for all the measures. On the other hand, comparing the Apicultur API and the Hybrid method implementation, the Hybrid method shows a better accuracy. This is due mainly for the performance in classifying neutral tweets; while in the case of classifying positive and negative tweets, the results vary.
6.2 Evaluation

The proposed solution has been developed as planned. The completion of the working products has been possible through the implementation of an agile developing methodology and the use of suitable tools.

This way, the proposed solution has integrated Java applications and R scripts to automatize the text processing tasks, and facilitate its execution as many times as required.

The design of the proposed solution presents strengths as well as weaknesses. Among the strengths, the following can be mentioned: modularity, simplicity and easy maintenance. On the other hand, the identified weaknesses include the implementation of the modules as independent applications that require integration, and the hard-coding of some parameters.

The present solution is still deployed on a development environment; therefore, if the system was required to be used in a real environment the following changes can be suggested:

- To create an enterprise application that integrates the developed modules.
- Parameters as collection names and file paths should be read from a configuration file or database, or alternatively an interface to provide the user the means to enter such values.
- To include extraction and storage of Twitter streaming data.
- To allow users accessing the sentiment analysis features through a web service.

The Web application that has been developed offers a simple, clean and easy to use interface for data visualisations. The featured graphs favour the results communications to users. Besides, the Primefaces component library utilised by the application, offers a wide range of chart components that can be easily invoked in the case that more visualisations are required. On the other hand, the Web application could offer an interface for maintaining the relational database tables, which store the decisions catalogue. Thus, the application only allows retrieving information at the moment.

The application utilised to read and store the tweets from a text file, accomplish this requirement satisfactorily. However, this process requires that the tweets are placed in the
file following a specific disposition or format. This implies extra efforts at the time of generating such files.

The pre-processing, sentiment feature generator and machine learning modules execute R functions that are called through the Java/R interface. This way, any change in the current process can be made by directly editing the code in R-scripts, without modifying the Java application code. In addition, the use of text files for holding intermediate results, help to reduce the load on the database. However, this also adds failure points at execution time, for instance, if a file is missing or cannot be opened.

The use of an out-of-the-box tool, in this case the online Apicultur API tool, have allowed developing a basic application for utilising other similar services, which in its majority are invoked through web services.

The application for classifying the tweets manually helped to carry out a laborious task. However, this basic application could be even more helpful if more features as filters or search options had been added to facilitate the classification.

Regarding the results obtained by the sentiment analysis method, the performance of the method did not achieve the high accuracy obtained by the method’s authors (>80%). However, this does not indicate that the selected method is not valid; it just demonstrates the well-known sentiment analysis limitations that have not been solved.

Regarding the learning experience, the entire process has been very enriching and challenging.

First, the time management was a serious concerning. The necessary number of hours per day that should be covered and the number of effective working hours were calculated in order to find a balance between them and carry out the work efficiently. Then, it was necessary to ensure an adequate study environment that maximise those effective hours.

The planning and the selection processes were critical for the accomplishment of the project. As the background research took more time than planned, task rescheduling and parallel execution considerations were necessary. Likewise, the research and agile development methods marked the path to follow in order to deliver a working solution. In addition, as the tools were thoroughly selected, there were no unbeatable obstacles that put at risk the conclusion of the project.

The background research and literature review elaboration were demanding tasks. At the beginning it was difficult to establish a way to carry out the study and select the appropriate resources. Later it was difficult to terminate the task due the amount of resources on the topic.

Although the proposed web application was simple, user centred techniques were utilised. It was good to proceed this way, as the analysis and design outcomes facilitated the implementation. Finally, the implementation was a satisfactory, motivating and extremely rewarding process.
7 Conclusions and Future Work

The following conclusions have been obtained:

- Regarding the selected tools, techniques and methods, the results have been satisfactory. The tools utilised for the solution implementation have met the expectations. Particularly, the use of MongoDB speeded up the text processing tasks by allowing the storing of new pieces of data as required, due to its schema flexibility features.

- The implemented hybrid approach achieved a lower accuracy than the results obtained by (Bahrainian & Dengel, 2013). And what is more important, the accuracy obtained for each decision dataset varies considerably; from a 78% in 2011 Referendum case, to a 65% in Assange’s asylum case and to a 35% in Yasuni exploitation case. This clearly indicates that the performance is subject to the context and data domain.

- The use of an already defined Spanish lexicon avoided the task of translating the tweets from Spanish to English. However, the lexicon only included a positive/negative classification of words and phrases, while the SentiStrength English lexicon provides a scale of values for classifying the words for its polarity as well as its intensity. Consequently, this also caused a negative impact on the method’s performance.

- The text processing steps added for the modified hybrid method did not improved the overall performance of the method but rather undermine it in some cases. This may indicate that the phrases lexicon replacement task diminished the effect of the stop-words removal, and that in some cases the stop-words should not be removed.

- The results presented through the Web application has allowed obtaining insights easily for both purposes, identifying the Twitter opinion on the particular decisions and comparing the assessment results of the methods utilised for this project. In addition, in the case of Yasuni exploitation decision, it was verified that peaks on the number of tweets coincides with the occurrence of particular events, by observing the corresponding linear graph.

- Through the manual classification task it was possible to identify that a great number of opinions are shared via videos, images or links, which prevents the opinion extraction from the tweet content. In addition, the use of hashtags also complicates the extraction of opinions as words cannot be easily identified and in some cases the entire hashtag may imply an opinion.

Presentation of word clouds for each decision dataset via the Web application, lexicon improvement by the insertion of specific domain words and the inclusion of a pre-processing clustering task for separating neutral tweets, are suggestions for future works.
8 References


