

**Decision Modelling Driven by Twitter Data: A Case Study of the 2017
Presidential Election in Ecuador**

**A thesis submitted to The University of Manchester for the degree of
Doctor of Philosophy
in the Faculty of Humanities**

2019

Lucia Bernarda Rivadeneira Barreiro

Alliance Manchester Business School

List of contents

Abstract.....	10
Declaration.....	11
Copyright statement.....	12
Dedicatory.....	13
Acknowledgments.....	14
Preface.....	15
Chapter 1 – Introduction.....	17
1.1. Research aims and research questions.....	18
1.2. Research generalisation and contributions.....	19
1.3. Thesis structure.....	19
1.4. Conference presentations and planned publications arising from this thesis.....	20
Chapter 2 – Research methodology.....	22
2.1. Type of research.....	22
2.2. Data sampling and analysis.....	23
Chapter 3 – Literature review.....	25
3.1. Value of Twitter data.....	25
3.2. Twitter-based sentiment model as a tool for measuring and anticipating future behaviour.....	27
3.3. Twitter-based predictive models to provide further understanding of user preferences.....	28
Chapter 4 – Decision behaviour analysis using Twitter data: A literature review.....	30
Abstract.....	30
4.1. Introduction.....	30
4.2. Criteria for literature search.....	34
4.3. Uses of Twitter data for analysing user behaviour.....	35
4.3.1. Effects of Twitter on users’ choices.....	36
4.3.2. Uses of Twitter as source of data.....	38
4.3.3. Twitter in politics.....	39
4.3.4. What comes after analysis of Twitter data?.....	41
4.4. Unpacking Twitter analysis.....	43

4.4.1. Sampling Twitter.....	43
4.4.2. Pre-processing.....	45
4.4.3. What analytical tools are recommended to meet the aims?.....	46
4.5. Challenges and critics of the use of Twitter data.....	49
4.6. Conclusion.....	53
4.7. Further research path.....	54
Appendix A-4.....	56
References.....	65
Chapter 5 – Making progress on Twitter data analysis: A study based on sentiment analysis and identification of influential users.....	79
Abstract.....	79
5.1. Introduction.....	79
5.2. The Twitter environment.....	83
5.3. Related work.....	86
5.3.1. Building models through sentiment analysis tools.....	87
5.3.1.1. Methods for conducting sentiment analysis.....	91
5.3.1.2. Performance evaluation of sentiment analysis with real cases....	92
5.3.2. Who is setting the agenda in political aspects?.....	93
5.3.3. Contribution of this paper.....	96
5.4. Methodology.....	96
5.4.1. Case study: The 2017 Presidential election in Ecuador.....	96
5.4.2. General information about Ecuadorian elections.....	97
5.4.3. Data sampling.....	98
5.4.4. Data pre-processing.....	102
5.4.5. Sentiment analysis.....	106
5.4.6. Identification of influential users.....	109
5.5. Results and discussion.....	113
5.5.1. Sentiment analysis results.....	113
5.5.2. Identification of influential users.....	118
5.5.3. Evolution of followers during the campaign.....	123
5.6. Conclusion.....	124
Appendix A-5.....	128
References.....	132
Chapter 6 – Predicting tweet impact using the evidential reasoning rule.....	145

Abstract.....	145
6.1. Introduction.....	145
6.2. Brief introduction to the evidential reasoning rule.....	149
6.2.1. The MAKER framework.....	150
6.2.2. The RIMER framework.....	152
6.3. Related work.....	154
6.3.1. Predictive models using Twitter data: Retweet analysis.....	154
6.3.2. Predicting retweets.....	155
6.3.3. Contribution of this paper.....	156
6.4. Methodology.....	159
6.4.1. Case study.....	159
6.4.2. Data sampling.....	159
6.4.3. Data analysis.....	161
6.4.3.1. Description of the model: Output.....	162
6.4.3.2. Description of the model: Inputs.....	163
6.5. Application of the ER rule to predict impact of tweets based on the number of retweets for the 2017 Ecuadorian Presidential election.....	169
6.5.1. Implementing the MAKER framework.....	170
6.5.2. Implementing the RIMER framework.....	174
6.5.3. Training the parameters of MAKER and RIMER.....	177
6.5.4. Validation with other machine learning approaches.....	178
6.6. Results and discussion.....	178
6.7. Conclusion.....	186
Appendix A-6.....	188
References.....	199
Chapter 7 – Conclusion and future research path.....	207
7.1. Towards a novel approach to Twitter analysis.....	207
7.2. Contributions and implication.....	210
7.3. Reflections on strengths and limitations.....	211
7.4. Directions for future research.....	211
References.....	213
Final word count: 44,980 words	

List of Tables

Table 4.1. Literature reviewed on the use of Twitter to develop predictive models.....	35
Table A-4.1. Summary of studies using Twitter data for classification-predictive modelling.....	56
Table A-4.2. Summary of studies that use sentiment analysis of Twitter data about politics.....	60
Table 5.1. Related work about Twitter feature pre-processing when performing sentiment analysis.....	89
Table 5.2. Appearance of the tweet in its original format after retrieving it through the Twitter Search API.....	101
Table 5.3. Example of emoticons and emoji coding.....	104
Table 5.4. Identification of tweets based on the source of the tweet. For illustration purposes, texts of tweets were translated from Spanish to English...	112
Table 5.5. Numbers of positive tweets and positive Twitter users about the 2017 Ecuadorian Presidential election before and after the pre-processing stage.....	113
Table 5.6. Twitter and official results from the 2017 Ecuadorian Presidential election.....	114
Table 5.7. Pre-vote polls and official results from the 2017 Ecuadorian Presidential election.....	118
Table 5.8. Proportion of tweets generated by different Twitter users regarding the two candidates.....	119
Table 5.9. Most influential users from those tweeting about Lenin Moreno.....	121
Table 5.10. Most influential users from those tweeting about Guillermo Lasso.	122
Table 6.1. Summary of models aiming to predict retweeting behaviour.....	158
Table 6.2. Total number of tweets generated by the two candidates, total number of retweets generated by other Twitter users during the elections, and descriptive statistics of retweets.....	160
Table 6.3. Estimates and normalised likelihoods for the first partial MAKER model of Guillermo Lasso comprising the variables type and emotion.....	172
Table 6.4. Interdependence index applied for the first partial MAKER model of Guillermo Lasso.....	173

Table 6.5. First partial MAKER model results after training weights of variables (type and emotion) and their parameters for Guillermo Lasso.....	174
Table 6.6. Illustration of the possible belief rules and belief degrees used for this case study.....	176
Table 6.7. Rule base using RIMER with updated belief degrees considering MAKER 1 and MAKER 2 partial model results for Lenin Moreno.....	179
Table 6.8. Rule base using RIMER with updated belief degrees considering MAKER 1 and MAKER 2 partial model results for Guillermo Lasso.....	180
Table 6.9. Comparison of performance of machine learning methods based on the MCE.....	181
Table A-6.1. Abbreviations of variables and their parameters used in this study.....	196
Table A-6.2. MAKER 1: Partial model involving emotion and URL for Lenin Moreno.....	197
Table A-6.3. MAKER 2: Partial model involving type, hashtag, and timeline for Lenin Moreno.....	197
Table A-6.4. MAKER 2: Partial model involving URL, hashtag, and timeline for Guillermo Lasso.....	198

List of Figures

Figure 4.1. Framework for modelling Twitter data to manage and tailor online campaigns.....	33
Figure 5.1. Identification of the interactive features that can be found in tweets.....	84
Figure 5.2. Electorate's age distribution during the second round of elections, 2017.....	98
Figure 5.3. Appearance of a tweet from Guillermo Lasso.....	100
Figure 5.4. Frequency plot of URLs found in tweets about Lenin Moreno.....	105
Figure 5.5. Frequency plot of URLs found in tweets about Guillermo Lasso....	106
Figure 5.6. Configuration of MeaningCloud™ for sentiment analysis purposes. Input allows the analysis of 10,000 tweets each time taking approximately 45 to 55 minutes to process the data.....	108
Figure 5.7. Steps taken for the sentiment analysis process.....	109
Figure 5.8. Evolution of sentiment on Twitter of the candidate Lenin Moreno during the two phases of elections.....	116
Figure 5.9. Evolution of sentiment on Twitter of the candidate Guillermo Lasso during the two phases of elections.....	117
Figure 5.10. Evolution of followers of the two candidates during the presidential campaign.....	124
Figure 5.11. Correlation plots between number of retweets and followers for both candidates.....	124
Figure A-5.1. Weekly breakdown of the frequency plot of URLs found in tweets about Lenin Moreno.....	129
Figure A-5.2. Weekly breakdown of the frequency plot of URLs found in tweets about Guillermo Lasso.....	131
Figure 6.1. Evolution of the number of retweets for both candidates during the sixteen weeks of campaigning.....	162
Figure 6.2. Original model developed to predict the impact of a tweet based on the number of retweets.....	164
Figure 6.3. Frequency plot of hashtags found in tweets about Lenin Moreno...	168
Figure 6.4. Frequency plot of hashtags found in tweets about Guillermo Lasso.....	168

Figure 6.5. Candidates' tweets per week.....	169
Figure 6.6. Hierarchical structures of the models for both candidates.....	175
Figure 6.7. Illustration of the MAKER-RIMER generic training process.....	177
Figure 6.8. Characteristics that make a tweet of high or low impact for Lenin Moreno.....	183
Figure 6.9. Characteristics that make a tweet of high or low impact for Guillermo Lasso.....	184
Figure A-6.1. Weekly breakdown of the frequency plot of hashtags found in tweets about Lenin Moreno.....	190
Figure A-6.2. Weekly frequency plot of hashtags found in tweets about Guillermo Lasso.....	192
Figure A-6.3. Weekly frequency plot of hashtags for Lenin Moreno during the campaign.....	193
Figure A-6.4. Weekly frequency plot of hashtags for Guillermo Lasso during the campaign.....	194
Figure A-6.5. Examples of tweets with high impact for Lenin Moreno and Guillermo Lasso.....	195

List of abbreviations

API	Application programming interface
BOW	Bag-of-words
BRB	Belief rule base
CNE	Electoral system authority of Ecuador
DJIA	Down Jones industrial average
DM	Direct message
D-S	Dempster-Shafer
DT	Decision tree
EEA	European Economic Area
ER	Evidential reasoning
EU	European Union
eWOM	Electronic word-of-mouth
FN	False negative
FP	False positive
GDPR	General Data Protection Regulation
GIF	Graphic interchange format
LR	Logistic regression
LIWC	Linguistic inquiry and word count
MAKER	Maximum likelihood evidential reasoning
MCE	Misclassification error
NA	Not available
NB	Naïve Bayes
NLP	Natural language processing
POS	Part-of-speech
RIMER	Belief rule-based inference methodology
SNA	Social networking analysis
SVM	Support vector machine
TF-IDF	Term frequency - inverse document frequency
TN	True negative
TP	True positive
UCG	User-generated content
URL	Uniform resource locator

Abstract

This thesis introduces novel approaches for using Twitter data for building models aiming to analyse decision behaviours in the political arena. The results, presented in the form of three academic papers, apply to problems of sentiment classification and machine learning approaches used for prediction tasks.

The first paper reviews the literature on the use of Twitter as a tool to analyse political behaviour. Particular attention is paid to approaches of user behaviour analysis, anticipation of outcomes, and predictive models. The paper identifies unresolved issues related to data selection and adequacy that can limit the performance of Twitter-based models, which researchers and practitioners, such as political campaigners, have not addressed in depth. In this regard, improvements in sampling, data pre-processing, and data analysis are likely to enhance the understanding of user behaviour in the political context. A practical implication, especially for campaigners, is the use of Twitter-based evidence to tailor communication strategies to entice different target audiences simultaneously, while also monitoring, measuring, managing, and evaluating the performance of campaigns.

The second paper introduces two novel approaches to Twitter analysis intended for enhancing the performance of sentiment analysis models and identifying influential users during electoral campaigns. For the former, a novel approach is proposed to pre-process Twitter data for sentiment analysis, which considers features in tweets, namely hashtags, emoticons, and URLs, that have been often discarded or not fully utilised in previous works. As for the latter, a new approach is proposed to identify influential users and their sentiment periodically using a two-stage process. A case study of the 2017 Ecuadorian Presidential election was used to develop and validate these approaches, in which 1.3 million tweets pertinent to the two most voted candidates were retrieved for analysis. The key findings are: first, the pre-processing approach improves sentiment analysis results in comparison to results using raw tweets. Second, the most frequent type of tweets observed are retweets, and the most retweeted content is often produced by the two candidates themselves. Third, the number of unique Twitter user accounts producing positive sentiment towards the candidates can provide a measure of vote share. In this study, the latter actually outperformed the results made by the officially authorised polling firms. These findings have implications for political marketing communication strategies that relate to identify sentiment of users towards candidates and influential users throughout a campaign on Twitter.

Finally, the third paper proposes a novel prediction model based on the evidential reasoning (ER) rule, named MAKER-RIMER, to predict whether the impact of a tweet is high or low in terms of the number of retweets it can achieve. The study relies on tweets produced by the two most voted candidates of the 2017 Ecuadorian Presidential election and uses five features of tweets as predictors. The proposed MAKER-RIMER model delivered an interpretable, transparent, and trackable model. Similarly, MAKER-RIMER performed better in terms of misclassification errors when compared against alternative machine learning prediction models. Last, this study identifies which features of tweets are causing impact of tweets to be high or low for each candidate. These findings support the design of Twitter content creation based on what users find more attractive to be retweeted.

Keywords: Twitter, Decision behaviour, Modelling, Sentiment analysis, Retweeting prediction, Evidential reasoning rule.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

- i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.
- ii. Copies of this thesis, either in full or in extracts, and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.
- iii. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.
- iv. Further information on the conditions under which disclosure, publication and commercialization of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see <http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420>), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see <http://www.library.manchester.ac.uk/about/regulations/>), and in The University’s policy on Presentation of Theses.

Dedictory

To my mom who after taught me how to walk she encouraged me to fly.
To my husband whose love is the most amazing gift I could even ask for.
To my two little children for being my inspiration to be better and do better.

Acknowledgments

I believe this is the most important part of the thesis because without the following people and institutions, I would not be able to complete this programme.

I offer my gratitude and appreciation to my supervisors, Prof Jian-Bo Yang and Dr Manuel López-Ibáñez. You have not been only my mentors, but a source of inspiration to challenge myself, to develop my critical thinking, and to shape my research abilities. I could not thank you more for sharing your knowledge with me, because without your guidance, this work would not have materialized. I would like also to extend my gratitude to the members of the Decision and Cognitive Sciences Research Centre (DCSRC) at the Alliance Manchester Business School for giving me the opportunity to discuss this work and getting critical feedback to improve it.

For the financial support, I would like to thank the National Secretary of Higher Education, Science, and Technology, SENESCYT, for my PhD scholarship. I want to extend my appreciation to MeaningCloud™ for providing me free access to use its software that supported part of the data analysis of this thesis. I am also grateful to the online R community for the technical assistance during these years. Exchange ideas and codes were vital for the development of this research.

A special thanks to my mom. Thank you for your patience and love during all my life, and for always encouraging me to find my potential and pursuing it. You are my greatest role model, and this thesis is your achievement as well. To the memory of my father, and the support of my siblings Julio, Daniel, Ma Piedad, and Eliana. To my parents in law, Wagner and Mariana, all my gratitude to you. You are the materialization of what families are supposed to be. I am also grateful for the amazing friends I have made during this time at the University of Manchester, without whom this journey would not be the same.

This thesis work is specially dedicated to my lovely husband, Ignacio, who has also shared his PhD journey with me. You have been my main source of inspiration, sanity, and support during the toughest times. I infinitely admire and love you, and I honestly believe I would not be able to complete this work without you. Finally, my gratitude and love go beyond to my two little children, Ignacio and Sofia. I hope you have enjoyed this PhD journey with us. Always remember you can fly as high and far as you wish. Go and spread your wings.

Preface

About five years ago, my master's thesis at Nanyang Technological University in Singapore was concerned with the willingness of online buyers to trust authentic and fake hotel reviews posted on the website of a leading online travel service provider (Expedia). Realising that the practice of posting reviews was in part a mechanism for misleading and influencing buyers' decisions motivated this work. Before conducting the master's thesis, I had little awareness that the information we consume on Internet could be used by others to understand, analyse, and model patterns of behaviour to subsequently implement mechanisms that influence users' perceptions and purchasing choices. Although my previous interest in fake news has refocused for my later academic career, my master's thesis was a point of inflexion to pursue a PhD that focuses on Twitter data and the possibility to use it as a tool to analyse user behaviours.

During my first year of the PhD, I had initially thought of the Ecuadorian inheritance law as a case study, for which public debate began in 2015, and which was finally approved in June 2016. This case study was intended to validate a model to analyse how bias could influence public reaction in social media users. However, the scope of that case study seemed rather narrow for generalisability. Also, by the time of data extraction, it was somewhat late to begin retrieving the tweets, which constituted an early limitation for using this case study. Because of these aspects, I chose a new case study for developing and validating models to analyse voting behaviour. I then refocused my attention toward the 2017 Ecuadorian Presidential election, in which the bias factor was no longer considered. This made an appropriate case study for different reasons. First, before starting the data sampling process I spent several months to define the purpose of the proposed models, and also, I had time to observe and discover eventual predictor and outcome variables arising in the 2016 United States Presidential election, which could be useful for studying the Ecuadorian election. Second, there were not many studies focusing on the analysis of user behaviour using Twitter in South America, so it was an opportunity to explore voting behaviour in the Latin American context in the Spanish language. Finally, no rivalry produced more Twitter content in Ecuador than this election, and the public opinion was highly polarised during those days. Hence, there was an opportunity to develop and validate decision models in the voting behaviour context.

Moreover, this thesis reflects my interest in social media's ability to mobilise sentiment and action in a particular direction, and with respect to a wide range of objectives. However, while the implications of this work on political management can be promising, I am also hoping that my contribution may support the mobilisation of action towards current global concerns such as sustainable development and climate change. Yet, the literature on the application of social media for analysis of user behaviour is more mature in the political field than elsewhere. This constitutes the best possible context and platform to learn from and advance the intended knowledge.

Finally, the rationale behind choosing to write this dissertation in the alternative format of three journal papers emerged since this study presents three major independent contributions to my area of research. In addition, I found this thesis format to be more suitable for my future career goals in Academia, since it has allowed me to familiarise myself with the process of academic publishing, including styles and format.

Chapter 1

Introduction

This thesis sets out to better understand how Twitter can analyse decision behaviour during electoral campaigns and to develop novel approaches for modelling purposes. Twitter is an open microblogging social media platform that allows users to share posts of no more than 280 characters called tweets. Created in 2006, Twitter has reached today 321 million monthly active users worldwide¹. Because of its popularity, Twitter has become a widespread tool for communication and dissemination of information (Clement, 2019), which is of special value for the media, politicians, and researchers. The study builds upon previous research on Twitter analytics for political purposes and uses Twitter data obtained during the official campaign period of the 2017 Ecuadorian Presidential election to develop two models. In doing so, this thesis sets forth new approaches based on Twitter data for enhancing sentiment analysis tasks, identifying influential users pertinent to the sentiment conveyed towards electoral candidates, and understanding what makes a candidate's tweet to achieve a high or low impact in terms of its number of retweets. These methods can contribute to understand decision behaviour during electoral campaigns.

A key issue in the Twitter literature that this thesis addresses is that its progress has been greatly based on intuitive rather than rigorous statistical examination. For example, although sentiment analysis² has been studied extensively and is today one of the most popular methods for Twitter-based models, there is not yet a set of criteria and general rules to conduct Twitter data sampling and to analyse tweet features such as hashtags, emoticons, and URLs. Similarly, retweets³ have been used widely as a proxy for popularity of a topic or Twitter user account. However, when it comes to anticipating the ability of a tweet to be retweeted, much of the literature has focused its attention on metrics of followership⁴, while the actual content value of tweets has received scant attention. On this matter, a set of guiding principles on how to anticipate the retweetability of content remains to be made.

¹ <https://www.washingtonpost.com/technology/2019/02/07/twitter-reveals-its-daily-active-user-numbers-first-time/?noredirect=on>

² Sentiment analysis is used to categorise opinions or feelings found in text, which can be labelled into positive, negative, and neutral sentiments (Nakov et al., 2016).

³ Retweet is the process of sharing or repost a tweet.

⁴ Followership refers to the number of followers a Twitter user has.

Moreover, machine learning approaches such as logistic regression, Naïve Bayes, decision tree, and support vector machine have been used for classification and modelling purposes in a variety of areas and using different data sources, including Twitter. However, while these methods have produced good results in terms of predicting, they present weaknesses in terms of transparency, interpretability, and trackability of the actions being suggested. These are all important considerations because politicians and campaigners benefit greatly from knowing why results occur. The above-mentioned are problems that this thesis will investigate and try to solve in the form of three academic papers. The remaining part of this introduction outlines the research aims and questions, research generalisation and contributions, a brief overview of the remaining chapters, and a summary of conferences and planned publications arising from this thesis.

1.1. Research aims and research questions

This thesis has three overarching aims. First, to provide an in-depth understanding of the ability of Twitter data to create models that can be used to analyse decision behaviour. Second, to improve approaches to Twitter-based sentiment analysis results and identification of influential users during electoral campaigns. And third, to make progress in the usability of Twitter-based machine learning predictive models, by focusing on aspects of transparency, interpretability, and trackability. The research questions addressing these aims are:

1. Where are we now and what comes next on Twitter data modelling relevant for the analysis of user decision behaviour?
2. How can data pre-processing help enhance Twitter sentiment classification accuracy and computational efficiency?
3. How can influential users that affect the Twitter sentiment be identified, and why is it important?
4. How can Twitter data be modelled so as to predict the retweetability of a tweet in a transparent, interpretable, and trackable manner?

Furthermore, the objectives behind these research questions are:

1. Identify the current state of knowledge about Twitter use for analysing decision behaviour in the political context.
2. Lay the groundwork for future research into improved methods for Twitter analytics.

3. Propose a data pre-processing approach to incorporate hashtags, emoticons, and URLs on Twitter sentiment analysis.
4. Propose a two-stage approach to identify influential Twitter user accounts that affect the sentiment towards a given Twitter user.
5. Propose a transparent, interpretable, and trackable machine learning model, based on the evidential reasoning (ER) rule, to predict the retweetability of a tweet.

1.2. Research generalisation and contributions

Although the scope of this thesis is limited to the Twitter setting, its intended innovation is actually the development of new approaches to support *Big data* modelling that are relevant for the analysis of decision behaviour.

Overall, this thesis contributes both to the understanding of Twitter as source of data for modelling purposes and to the development of predictive models for improved data interpretation. Specifically, this thesis makes the following contributions:

1. The identification of gaps in the literature and research opportunities for improving the performance of Twitter data modelling for the analysis of decision behaviour.
2. A data pre-processing approach that incorporates information contained in hashtags, emoticons, and URLs, when modelling Twitter data, intended for improved accuracy and computational efficiency when conducting sentiment analysis.
3. An approach to periodically identify influential Twitter user accounts that influence sentiment polarity.
4. A predictive model based on the ER rule to estimate the impact of a tweet, in terms of retweets, in a transparent, interpretable, and trackable way.

1.3. Thesis structure

The remainder of this thesis is structured as follows: Chapter 2 provides a brief overview of the methodology used to address the research questions and objectives. Chapter 3 presents an introduction to the state of the literature to set the tone for this study. Then, the three academic papers are presented in the subsequent chapters as follows. Chapter 4 reviews the literature on the use of Twitter-based models for decision behaviour analysis in the political arena. In doing so, it identifies

areas for future research, which are relevant for the performance of Twitter-based models used for user behaviour analysis.

Chapter 5 presents the second paper. It builds on Twitter data pertinent to the 2017 Ecuadorian Presidential election. The paper develops two approaches intended to enhance the performance of sentiment analysis models. First, an approach to pre-process tweets intended for sentiment analysis. Unlike previous studies, this approach considers hashtags, emoticons, and URLs. And second, a two-stage approach to periodically identify influential users and their sentiment towards the two most voted candidates.

Chapter 6 presents the third paper. Using the same case study as in the previous chapter, it develops a predictive model based on the ER rule to predict the impact of a tweet. Although progress in prediction accuracy is attempted, this model primarily aims to be more transparent, interpretable, and trackable than alternative machine learning approaches. The ER based model, called MAKER-RIMER, where MAKER stands for maximum likelihood evidential reasoning, and RIMER for belief rule-based inference methodology using the ER approach, codes five features of tweets produced by the two most voted candidates of the 2017 Ecuadorian Presidential election. This approach identifies which features drive tweets to be of high or low impact for each candidate.

Finally, the conclusion of this research is presented in Chapter 7, which summarises and highlights the main arguments and contributions of this study, acknowledges its strengths and limitations, and provides directions for future work.

1.4. Conference presentations and planned publications arising from this thesis

The contributions of this thesis have been introduced in several conferences, and are being prepared for publication in decision science journals:

- Annual Conference on Engineering and Information Technology
ACEAIT, 2019 (Kyoto, Japan): Refereed conference paper presenting the results of the sentiment prediction as shown in Chapter 5 of this thesis.
- Operational research OR60, 2018 (Lancaster, United Kingdom): Refereed conference abstract concerning the results of the evidential reasoning (ER) prediction as shown in Chapter 6 of this thesis.
- Advances in Data Science, 2018 (Manchester, United Kingdom): Poster presentation of the results of the ER prediction as presented in Chapter 6.

- International Conference on the City IAFOR, 2017 (Barcelona, Spain):
Refereed conference abstract presenting an overview of the methodology
of the ER prediction shown in Chapter 6.
- PGR Conference 2016 (Manchester, United Kingdom): Overview of main
aspect of the thesis.
- Planned publication of Chapter 4 in the journal Human-centric Computing
and Information Systems.
- Planned publication of Chapter 5 in International Journal of Information
Management.
- Planned publication of Chapter 6 in the journal Expert Systems with
Applications.

Chapter 2

Research methodology

This chapter presents the methodology and methods used for collecting and analysing the data in the thesis. It first discusses the type of research design adopted, followed by a brief introduction of the research methods, which will be covered in more details in the three academic papers presented in Chapters 4, 5, and 6.

2.1. Type of research

This thesis proposes new approaches to analyse Twitter data, which can assist academics, political campaigners, and politicians in understanding users' preferences and what influences their choices. Political campaigners, from now on called campaigners, refer to the people involved in electoral campaigns, who seek to influence vote choice.

The thesis begins by reviewing the literature on the use of Twitter as a research tool and identifies future research directions in this field. Then, it proposes a data pre-processing approach for enhancing sentiment analysis on Twitter, and an approach to identify influential users involved in conveying sentiment towards candidates during an electoral campaign. Finally, the thesis develops a predictive model to understand the features in tweets that lead to high or low impact in terms of the number of retweets for candidates of an electoral race.

The study employs a quantitative, inductive, and designed-based approach to studying Twitter data. It uses mathematical approaches to transform raw data into information that supports user behaviour analysis. Quantitative research is suitable to develop predictive models that explain behaviours and processes, and can serve as grounds to anticipate future results (Bertrand & Fransoo, 2002). The first model which is developed in Chapter 5, detects the sentiment of Twitter users and the influential users impacting it. The second model, developed in Chapter 6, provides a tool to understand what is influencing the decision of Twitter users to retweet. It presents a machine learning model that can help campaigners to enhance the retweetability of tweets. This model is then compared to other machine learning methods in terms of accuracy.

This thesis also employs an inductive approach to generate models based on patterns identified from the data. The inductive approach is suitable to explore new phenomena or to look at previous ones from different perspectives (Gabriel, 2013). In

addition, this work uses design-based research, which is a suitable approach for technology-enhanced learning (Hadjerrouit, 2008; Pibernik & Dolić, 2008). The design comprises a case study used to develop and test the sentiment and retweeting predictive models, which can create an impact on practical situations. For instance, during electoral campaigns, candidates could further the ability to adapt communications according to the group of users they want to reach, produce tweets that reach a larger audience, and engage influential users to broadcast key messages. While developing the models, the study registers the processes of data sampling, analysis, and testing. Finally, this research uses Twitter because it provides a dynamic environment to analyse real-world problems as they occur.

2.2. Data sampling and analysis

This thesis is structured into three academic papers, each using a different type and source of data. The first one reviews the literature on Twitter-based models used to analyse decision behaviour and identifies paths for future work. Several academic search engines were used to search the relevant literature; the criteria involved the keywords “Twitter”, “prediction”, “predictive analytics”, “predictive model”, “sentiment analysis”, and “decision behaviour”.

To build the models for the second and third papers, this thesis uses the case study of the 2017 Ecuadorian Presidential election. Approximately 1.3 million tweets were extracted using the Twitter Search API (application programming interface) and R Core Team (2013) during the two rounds of the official campaign period. Data extraction for the first round lasted 12 weeks, from November 2016 to February 2017, where a total of 8 candidates entered the race. The second round lasted 4 weeks, from March to April 2017 with the two most voted candidates, Lenin Moreno and Guillermo Lasso. Finally, Lenin Moreno was elected President of Ecuador. For the data pre-processing approach, as will be covered in Chapter 5, hashtags, emoticons, and URLs were coded into readable text. Then, sentiment analysis was performed with the text analytical software MeaningCloud™. As for the most influential users during the campaign, the five accounts obtaining the highest number of retweets and their sentiment towards candidates were identified every week.

The third paper uses the tweets that Lenin Moreno and Guillermo Lasso posted throughout the campaign. Its aim is to determine the impact of their tweets, being either high or low, based on the number of retweets. The variables used to develop the model are divided into output and inputs. The output is the number of

retweets achieved by every tweet. The inputs comprise categorical values for the type of tweet, emotion, URL, hashtag, and timeline. For comparison purposes, five machine learning methods were tested: logistic regression, Naïve Bayes, decision tree, support vector machine, and a proposed novel model based on the ER rule called MAKER-RIMER. By using this methodology, this chapter attempts to provide a better understanding of Twitter-based predictive models for retweeting purposes and addresses several gaps in the literature that are relevant for improving performance in terms of accuracy and transparency.

Chapter 3

Literature review

This chapter reviews the conceptual foundation for this thesis. Specifically, it investigates how Twitter-based models support the understanding and analysis of user behaviours during electoral campaigns. The perspective of this thesis is mainly that of politicians and campaigners as decision makers in the political context.

The past decade has seen the rapid development of Twitter-based research literature. During this timeframe, studies have focused on developing models to transform tweets into knowledge to understand and analyse decision behaviour, and on improving the ability of these models to provide accurate results. Yet, there are some aspects and features of Twitter that have not been fully addressed, which might have an influence on the performance of predictive models. In this sense, this chapter attempts to provide a brief overview of previous works and identify prominent gaps in the field.

This chapter first gives a brief overview of the value of Twitter data. The second part discusses the use of predictive models for anticipating outcomes. Finally, the third part includes a discussion of the use of Twitter to analyse retweeting behaviour.

3.1. Value of Twitter data

According to Isson (2018), 80% of *Big data* is unstructured, and only 0.05% of these data are analysed. Unstructured *Big data* refers to heterogeneous data usually found in text, audio, video format, and social media platforms, which lack the structural organisation that machines require for analysis (Gandomi & Haider, 2015). These data cannot be immediately used, but instead need to be processed and converted into a readable format for analysis. Companies that do not analyse unstructured data tend to miss competitive advantage compared to those actually doing so (Isson, 2018). In this sense, unstructured data found in social media platforms such as Twitter are fast becoming a key tool to understand and analyse user decision behaviour.

Twitter is a popular social media platform used by academics and campaigners in different fields to analyse users' behaviours. Twitter has been used as data source to assist human decisions, which in most cases rely on predictive analytics. In general, predictive analytics seek to discover patterns and capture relationships in the data, and

transform historical and current data into readable formats to support decisions and improve profitability (Gandomi & Haider, 2015; Waller & Fawcett, 2013). Moreover, predictive analytics encompasses the use of data mining and statistical techniques to make future predictions, and identifies risks and opportunities for political purposes (Schlegel, 2014). One of the tasks of predictive analytics involves the development of classification-predictive models, which focus on a mapping function aiming to predict the class or category of an output variable from a group of input variables (Brownlee, 2017).

Predictive model development is relevant for analysing decision behaviour, especially for campaigners of political parties, since it can help to anticipate how users might behave and to identify trends from the evidence of the data. This study focuses on two purposes behind the development of predictive models. One of these purposes is to enhance the ability of sentiment analysis models to measure and anticipate vote share using Twitter data. Today, campaigners use sentiment analysis widely to identify prevalent users' feelings and to address users' negative perceptions about a candidate (Ingle, Kante, Samak, & Kumari, 2015; Khatri & Srivastava, 2016). The second purpose is to use predictive models for a proactive management of Twitter user accounts in ways that enhance the ability to influence attitudes and behaviours of users. These models are relevant for campaigners since they can help design tweets based on the target audience.

However, the use of Twitter data for the two purposes mentioned above still presents some limitations. To name a few, criticism has focused on the limitations of existing data pre-processing approaches (Bao, Quan, Wang, & Ren, 2014; Thakkar & Patel, 2015), which will be addressed in Chapter 5 and Chapter 6, representativeness (Vijayaraghavan, Vosoughi, & Roy, 2017; Wright, Golder, Balkham, & McCambridge, 2019), and the spread of fake users and misleading content (Gupta, Lamba, & Kumaraguru, 2013). Fake users refer to Twitter user accounts created intentionally to manipulate users' perceptions, and often perform automatic or semi-automatic actions (Gurajala, White, Hudson, & Matthews, 2015). Misleading or fake content, on the other hand, is the spreading of "speculation, rumours, and mistrust" (DiFranzo & Gloria-Garcia, 2017, p. 38). These limitations constitute a basis for future research.

More specifically, Twitter aspects that can be explored to enhance the development of predictive models include, for example, sampling techniques. Much previous work has been conducted the sampling task intuitively. So, further research

could focus on establishing sampling guidelines in terms of the number of tweets retrieved and the timeline for data sampling. Also, there are features of tweets, such as hashtags, emoticons, and URLs, which could be included and analysed to enhance the performance of predictive models. Lastly, concerning fake users and misleading content, it should be noted that the literature review has focused on the identification and detection thereof. However, further analysis can be done to determine to what extent they can actually influence users' behaviours, or, as Shin, Jian, Driscoll, and Bar (2018) have argued, how they can represent a threat to democratic societies.

3.2. Twitter-based sentiment model as a tool for measuring and anticipating future behaviour

As explained in the previous section, one of the purposes for developing predictive models is to measure and anticipate future behaviour. On this matter, plenty of studies have used sentiment analysis approaches in areas such as politics (Ceron, Curini, & Iacus, 2015; Elghazaly, Mahmoud, & Hefny, 2016; Hürlimann et al., 2016), health (Alayba, Palade, England, & Iqbal, 2017; Jull et al., 2016), stock market (Oliveira, Cortez, & Areal, 2017; Pagolu, Reddy, Panda, & Majhi, 2016; Skuza & Romanowski, 2015), and in the sports field (Sinha, Dyer, Gimpel, & Smith, 2013; Yu & Wang, 2015). Some of these studies have focused on developing machine learning models, for which the use of corpora, which are datasets of texts containing sentiment polarity, is needed to determine whether tweets contain positive, neutral, or negative feeling connotation. In the absence of corpora, the use of automated sentiment detection software is an alternative for conducting sentiment analysis.

For the development of sentiment analysis models, a key task involves the data pre-processing stage, which implies converting unstructured text into a readable format. Before applying approaches to *clean* the data, it is necessary to select the features of tweets that have to be included for later analysis. However, features such as hashtags, emoticons, and URLs, have been often discarded from analysis, because it is believed that they contribute to noisiness and do not add value to the analytical and predictive stages (Al Hamoud, Alwehaibi, Roy, & Bikdash, 2018; Jain & Jain, 2019; Kumar & Babu, 2019; Sharma & Moh, 2016). Or, if they have been included, this has been often done simplistically by replacing them with tokens or applying quantitative codification (Gokulakrishnan et al., 2012; Pandarachalil, Sendhilkumar, & Mahalakshmi, 2015). Therefore, the informational value of these features has been overlooked in previous work. The inclusion and enhanced usability of these in

sentiment analysis models is necessary because they can provide a deeper understanding of user sentiments. Aiming to overcome this limitation, this thesis will cover a novel approach for data pre-processing which will be covered in Chapter 5, Subsection 5.4.4 using features of tweets that have been previously discarded from analysis.

Another practice used by academics and campaigners when aiming to understand user behaviour is the identification of influential users, who have the ability to influence others' behaviours (de Maertelaere, Li, & Berens, 2012). Traditionally, this influence has been measured in terms of the number of followers or retweets from an intuitive perspective (Montangero & Furini, 2015). Chapter 5 proposes a two-stage approach to understand how the Twitter network is distributed and identify the most influential users. Unlike previous research, this study defines influential users based on the data sources. Specifically, this study identifies the most frequent type of tweet used in the datasets to define what constitutes an influential user. Here, types of tweets are retweet, reply, and personal tweets. Since in the datasets used in this study the most common type of tweet is retweet, the definition of influential user is based on the number of retweets. The next step is to identify, periodically, the tweets that achieve the highest number of retweets and the users who have written them. These users, in the context of this study, constitute the influential users. Lastly, it is also important for campaigners to identify the sentiment position of the influential users towards specific Twitter user accounts, since it is likely that these users help shape the opinion of other users with regard to the candidates. The detection of sentiment of the most influential users towards the candidates will be covered in Subsection 5.4.6.

3.3. Twitter-based predictive models to provide further understanding of user preferences

Twitter-based predictive models, in the political arena, are also used to identify what needs to be addressed to gain the support of users. This is often measured in terms of changes in number of followers and in the number of retweets produced by a single tweet. Cha, Haddadi, Benevenuto, and Gummadi (2010) and Hong, Dan, and Davison (2011) support the view that the number of retweets is a proxy for influence, and depends on both the tweet content and name value (user producing the tweet).

Retweet predictive models have been developed from different perspectives. Previous studies have focused, for example, on predicting if a tweet will be retweeted (Hong et al., 2011; Nesi, Pantaleo, Paoli, & Zaza, 2018; Petrovic, Osborne, & Lavrenko, 2011), or on identifying motivations and elements of content more susceptible to being retweeted (Choi, 2014; Naveed, Gottron, Kunegis, & Alhadi, 2011; Trilling, Tolochko, & Burscher, 2017). To build the models, these studies have used Twitter features such as the number of URLs, hashtags, or followers, as predictor variables. Then, traditional machine learning methods such as logistic regressions, decision tree, or support vector machine, have been used for analysis and compared for suitability. However, these approaches have given little to no attention to the content values of such Twitter features. To address this issue, Chapter 6 proposes a prediction model based on the ER rule to determine the impact of tweets based on the number of retweets considering five features of tweets as predictors.

In summary, previous research has focused on studying and developing Twitter-based models aiming to support the understanding of user behaviour. However, there are areas still under-explored that can provide a better understanding of users' behaviours and more functional predictive models. This thesis attempts to address some of these gaps via the three academic papers presented in Chapters 4, 5, and 6, where a more detailed literature review specific to each of the papers will be provided.

Chapter 4

Decision behaviour analysis using Twitter data: A literature review

Abstract: This paper provides a systematic review of the literature on Twitter as data source for decision behaviour analysis in electoral campaigns. It has a twofold purpose: first, to understand how Twitter data have been modelled to analyse user interests and predict user behaviours; and second, to identify paths for future research. Emphasis is placed on approaches to measure and anticipate user behaviours with sentiment analysis approaches, and predictive models aimed at maximising goal-seeking tasks. The paper also identifies challenges of which academics and campaigners need to be aware when using Twitter as source of data. This study identifies pending issues related to data selection and adequacy that may be limiting the performance of Twitter-based models. Progress in these models may be produced if there is an improvement in sampling, pre-processing, and analysis that can support further understanding of users' behaviours in different contexts. As a result, campaigners can use Twitter-based evidence to develop suitable and personalised marketing strategies for the target audience, while being able to monitor, measure, manage, and evaluate the overall performance of the campaigns.

Keywords: Twitter, Literature review, Decision behaviour, Predictive modelling.

Declaration of interest: None.

4.1. Introduction

This paper reviews the literature on how researchers and political parties, with particular reference to campaigners during electoral contests, use Twitter to understand and analyse user choices and behaviours during electoral campaigns. Twitter is a microblogging site where posts are called tweets. The everyday dynamic on Twitter depends on following other users, so that it is possible to observe their tweets. Then, a network of followers is shaped, in which each user has followers (other users that can observe one's tweets) and followees (which make other users' tweets observable by a follower). Relationships and interaction within the followers' networks do not need to be reciprocal on Twitter.

Understanding how people make choices has been a subject of research for generations. For this case, choice means the cognitive process of selecting among two or more possible options. In a rational context, this would involve the choice taker

identifying the options at hand and assessing the outcomes for each possible decision. With the emergence of Twitter and other social media platforms, new light has been shed on the way people make choices. Of particular concern is the way information in social media influences stratification of preferences, and ultimately choices (Goodrich & De Mooij, 2014). This understanding can help campaigners to anticipate user behaviour and choices and thereby help politicians manage their Twitter user accounts in ways that influence potential voters. For example, Twitter analysis has been used to tailor content intended to increase awareness about candidates during a campaign (Pons, 2016; Schipper & Woo, 2017).

Traditionally, academics and campaigners have relied on different approaches for monitoring and forecasting the outcomes of their campaigning efforts. Surveys, questionnaires, interviews (conducted face-to-face, over the telephone, or via mail), and observations have been the dominant methods of data sampling. Moreover, despite the recent surge of Twitter-based analytics approaches, the traditional methods continue to be widely used. This is because traditional methods still have advantages that sources of Twitter, or other social media platforms, are not yet able to match (Mellon & Prosser, 2017; Sinnenberg et al., 2017). To mention a few: first, the traditional methods produce well-structured datasets, whereas those from social media are often unstructured and require more sophisticated analyses; second, sampling in traditional methods often makes it possible to obtain better targeted data for analysis; and third, traditional methods are more effective when discussing sensitive topics, particularly when investigating introverted societies. However, traditional methods also present some disadvantages, which are partially boosting the use of social media for similar purposes. These include: first, they are expensive to conduct and time-consuming; second, they are less effective in societies with restricted freedom of speech (Reuter & Szakonyi, 2015; Skoric et al., 2012); and third, participants in traditional methods are more prone to bias their answers (Kalimeri et al., 2019).

One of the main distinguishing features of Twitter, and other social media platforms, is that data are being continuously generated and consumed by users. It means that this type of source allows the extraction of both historical, and more importantly, real-time data. Moreover, this dynamic of sharing information, thoughts and attitudes often exposes users' demographic aspects, interests, and behaviours (Brena et al., 2019). In this sense, the understanding and monitoring of user behaviour by campaigners can be supported in a more robust fashion, as, for instance, they may track previous and current users' behaviours or identify consumption patterns.

As to methods used to analyse user behaviours, the application of predictive analytics stands out. This refers to the use of historical data, empirical records, or statistical evidence to generate models able to predict future outcomes, and to identify patterns that can assist exploratory modelling (Shmueli & Koppius, 2011). Through predictive analytics, Twitter content can be used to measure and evaluate the effectiveness of strategies and increase the competitive advantage over those competitors who do not use this source of data (Ibrahim, Wang, & Bourne, 2017). Thus, predictive analytics can support campaigners in designing their communication strategies based on real-time knowledge. For example, adapted from Pons (2016), a young woman who often retweets content of a family magazine would be, more likely, targeted to receive information on a candidate's proposal to improve childcare and schooling. This example suggests that Twitter-based predictive analytics can help capture demographic and interest data such as sex and age group and classify the content of tweets according to user interests. Therefore, interpretation of data can support personalisation of tweet content.

Figure 4.1 displays an elementary Twitter-based predictive analytics approach that can be used for managing Twitter user accounts towards objectives, such as higher vote shares and user support. Once the sampling criteria are set based on the context under study, data are extracted from Twitter using data mining techniques. Then, the behaviours of the users need to be initially observed through an exploratory analysis, followed by the analysis of the effects. Based on the preliminary results, and depending on the desired target, a strategy can be developed for its implementation using predictive models. However, since the production of the data on Twitter occurs in a continuous fashion, it is essential that campaigners continue observing users' behaviour to determine if the strategy needs to be re-adapted or optimised. This framework provides a generic tool that allows campaigners to make informed decisions based on evidence generated from data. If applied in real-time, this framework can help campaigners to make continuous predictions and timely evaluate the strategy implementation.

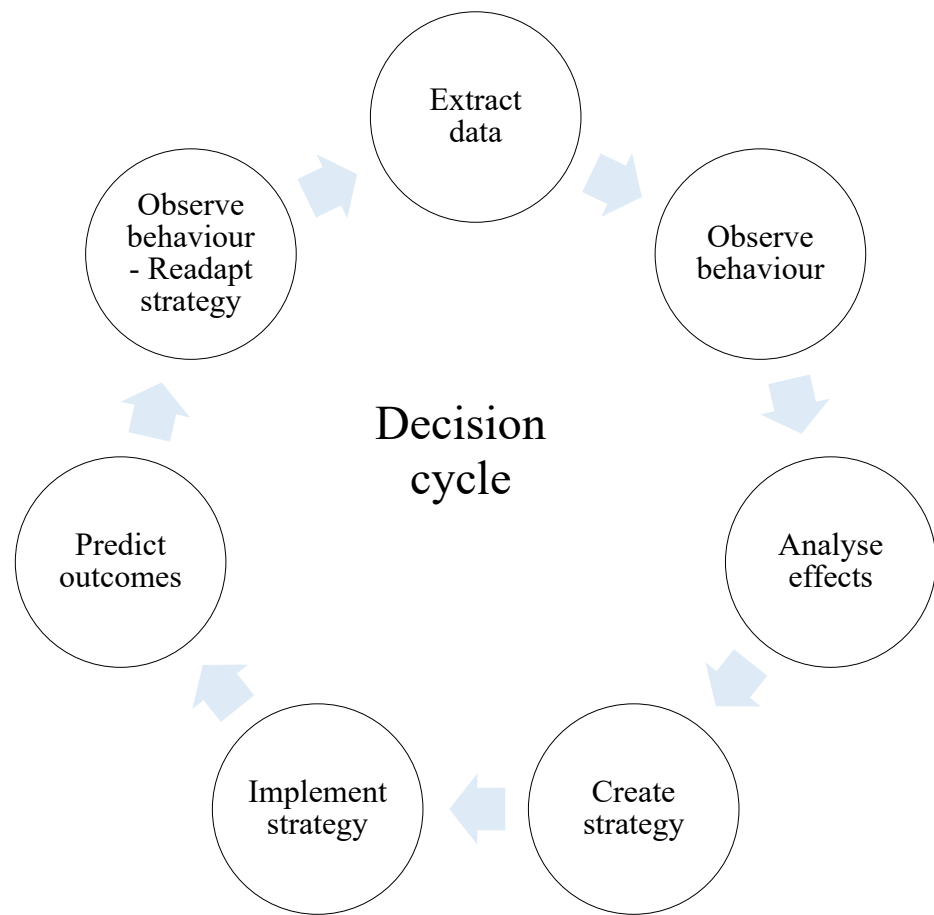


Figure 4.1. Framework for modelling Twitter data to manage and tailor online campaigns.

However, Twitter data present challenges and require caution for modelling purposes. As mentioned above, one challenge is that data can take the form of semi-structured data, meaning that they might or might not have predefined structures. Indeed, overcoming lack of structure is perhaps the main challenge for using Twitter as a source of data for monitoring and prediction purposes. Structured data are a type of well-schemed data, usually sorted in a database, e.g. measurements, signals, or indices. Twitter has some structured metrics such as the number of followers, favourites, and retweets, but most of the rich information comes in the form of freely written text, images, videos, and audio files (Martin-Sanchez & Verspoor, 2014), which are composed of natural language written by users and are unstructured by nature. This means that the unstructured data are not ready for use directly as inputs in modelling, but instead need to be interpreted and transformed into readable records for computing purposes. With the penetration of the Internet and use of social media

platforms for daily activities, it is expected that attention to analytics of unstructured data will continue to sharply increase through the years.

The aim of this paper is twofold. First, to reflect on the developments on Twitter in terms of how it has been used to analyse user behaviour for prediction purposes. This is done by discussing the current state of knowledge on Twitter as data source for academics and campaigners. The second aim is to detect areas for future study that can enhance the functionality of Twitter-based predictive models. Although there is a plethora of research about Twitter as data source for predictive modelling, the performance of predictions reported so far indicates that there is still vast room for future research, perhaps in the direction of gaining more meaningful insight from the data. The rest of this paper is structured as follows. Section 4.2 briefly describes the selection criteria used to retrieve the literature. Section 4.3 discusses Twitter as data source to analyse decision behaviour and reviews Twitter-based predictive models as reported in the literature. Section 4.4 unpacks the process of Twitter analysis as previously conducted. Section 4.5 identifies the challenges and criticisms associated with the use of Twitter as a source of data. The concluding remarks are presented in Section 4.6, and possible paths for future research are shown in Section 4.7.

4.2. Criteria for literature search

The use of Twitter as data source for predictive analytics purposes has been a growing subject of interest in the last decade. The Google Scholar database was initially searched for English-language literature by using the keywords “Twitter” and “prediction”, which returned more than 3.6 million results. When one includes the keyword “decision behaviour”, the search is reduced to approximately 450 thousand results, but this includes a large portion of healthcare-related publications that give limited attention to the actual development of the predictive models. To further filter these results, the keywords “Twitter”, “prediction”, “predictive analytics”, “politics”, “predictive model”, “sentiment analysis”, and “decision behaviour” were combined iteratively, resulting in 765 publications. Then, publications for review were selected using a two-filter criterion: first, publications that target Twitter sentiment analysis, Twitter predictive models, case studies on political elections, and challenges associated to the use of Twitter data; and second, documents in the format of academic papers, conference proceedings, and books available in the databases Springer Link, Scopus, IEEE Xplore, arXiv, and CiteSeer. Table 4.1 provides a summary of the literature selection reviewed for this study consisting of 226 studies.

Table 4.1

Literature reviewed on the use of Twitter to develop predictive models

Year	Studies using Twitter for predictive modelling		
	Journals	Proceedings	Books and reports
As of 2009	13	4	5
2010	5	10	1
2011	4	8	1
2012	7	5	2
2013	8	11	4
2014	19	5	5
2015	12	12	6
2016	6	10	3
2017	21	5	5
2018	12	1	7
2019	7	1	1
Total	114	72	40

4.3. Uses of Twitter data for analysing user behaviour

Academics and campaigners have used Twitter data to predict behaviour of a targeted population, and to gain knowledge that can help designing strategies and managing Twitter user accounts in ways that positively engage potential voters (Goodrich & De Mooij, 2014; Kaisler, Armour, Espinosa, & Money, 2013; McAfee et al., 2012; Sagioglu & Sinanc, 2013). This is achieved by transforming the data relevant to users' choices into knowledge, which is later used to make informed decisions that can influence the performance of a campaign.

In the business realm, the use of Twitter has proven effective for a number of companies to become more responsive and competitive (McAfee et al., 2012). Predictive modelling using Twitter has been used for a wide range of purposes, which include raising brand awareness, building brand engagement, and promoting positive eWOM (Hoffman & Fodor, 2010). eWOM (electronic word-of-mouth) refers to the online exchange of feedback about products and services, and is considered a way for users to influence others' decisions through exposure to social media platforms (Brown, Broderick, & Lee, 2007). It is also believed to be an important influence on revenue generation for companies (Chen & Shen, 2015; George, Haas, & Pentland, 2014). When Twitter campaigns are successful, users tend to be more loyal to brands

or actors, while becoming less responsive to the exposition of negative feedback (Culnan, McHugh, & Zubillaga, 2010).

In some way, users are progressively depending on Twitter not only to inform themselves, but also to support their choices, while fulfilling personal and professional needs (Blight, Ruppel, & Schoenbauer, 2017). People make choices on a continuous basis, from basic and routine choices to challenging ones that involve high degrees of uncertainty and occur under difficult situations. However, there are different and fluctuating features that impact choices across contexts, for which, understanding patterns of choice continues to be a major challenge to human behaviour academics.

4.3.1. Effects of Twitter on users' choices

The ways in which people make choices have been studied from different angles. Rolls et al. (2019) stated that there are two structures triggered in the brain that influence choices. The first refers to a system of reward and punishment, which cannot be simultaneously optimised. This system, which is associated with emotional and motivational behaviour, is often based on heuristics and intuition, where the choice taker eventually relies on automatic and unconscious processes, especially in contexts with a higher degree of uncertainty, or when there is not sufficient information to support a rational choice (Davies, 2015). The second involves the use of reasoning to make choices through syntactical thought. A limitation here is that the brain tends to be poor when conducting logical assessments, which can tend to irrational choices. In this sense, choices can be the result not only of logical thinking and statistical methods, but also of the power of experience, intuition, and mental simulation (Klein, 2017).

In the social media context, behavioural decision theory (Lau, 2003; Redlawsk, 2004) can help understanding the reasoning behind the way people make choices. This theory states that people make choices based on the information they have available. Moreover, since most users are limited information processors, they tend to pursue good choices with the minimum cognitive effort. Choices, consequently, are likely to be taken based on the perception of making a good choice, rather than upon estimating the value-maximising alternative. To some extent, cognitive limitations could lead users to choose a poor alternative. Thus, choices can be affected by the complexity of options and time pressure, as well as the

overwhelming and chaotic amount of information available on which to base choices (Redlawsk, 2004).

In the political context, factors that influence voting behaviour are well established, including the prominent ones: ideology (Ensley, 2007; Lachat, 2008), incentives and cognitive ability to make an accurate judgement (Davies, 2015; Ensley, 2007), emotional state of voters (Cwalina, Falkowski, & Newman, 2010), polls and media (Cwalina et al., 2010), access to more and privileged information (Brown, Hillegeist, & Lo, 2004; Davies, 2015; Lau, 2003), demographic aspects, political awareness (Schofield & Reeves, 2015), and prior beliefs (Zimper & Ludwig, 2009). These beliefs are conceived as the opinions and attitudes that users hold regarding a specific issue.

Moreover, several academics of political psychology claim that political information online, particularly the kind displayed on social media platforms, can also influence voting behaviour and political choices (Cottam, Mastors, Preston, & Dietz, 2015; DiGrazia, McKelvey, Bollen, & Rojas, 2013; Goodrich & De Mooij, 2014), especially in young voters (Munir, 2018; Sharma & Parma, 2016). Yet, there are also academics who maintain that voting choices are unaffected by social media and that, instead, online platforms support confirmation bias (Knobloch-Westerwick, Johnson, & Westerwick, 2014; Spohr, 2017; Zimper & Ludwig, 2009). Confirmation bias is manifest when users reject emerging information that may challenge their prior beliefs about politics. Additionally, social media users who discuss politics tend to be surrounded by other users with similar political stances. This tendency is known as social network homophily (McPherson, Smith-Lovin, & Cook, 2001). Users, therefore, are prone to create social connections with like-minded users, or with those sharing specific demographic characteristics, while isolating themselves from others with different viewpoints. As a result, they form biased beliefs that support their original posture on an issue, rather than taking into account emerging evidence that may challenge any such original stance (Poortinga & Pidgeon, 2004; Zimper & Ludwig, 2009).

Moreover, not only does access to value-relevant information influence political choices, either challenging or reinforcing them, but can also prompt users to share their views and take part in the conversation. On Twitter, this is manifest in the creation of pertinent content or when replying or retweeting, which implies support or opposition towards a candidate. Therefore, when creating Twitter content during an electoral campaign, besides influencing vote choice, it is important to anticipate how

users might interpret and use such content. In general, it seems that the analysis of decisions in fields such as voting behaviour is a complex process (Kazimieras Zavadskas, Antucheviciene, & Chatterjee, 2019; Redlawsk, 2004; Revelli, 2002; Rolls et al., 2019) and remains sometimes as a major challenge to understanding and predicting user behaviours.

Also, with the increasing prevalence of Twitter as a news source, concerns have arisen regarding the influence of fake users and misleading information in politics (Vosoughi, Roy, & Aral, 2018). The possible effects of misinformation on voting behaviour have been at the core of recent political scandals. Supposedly, political information has been manipulated to influence voting behaviour and favour specific candidates for either the use of “bot” Twitter user accounts, or by hacking of Twitter user accounts as claimed during the latest 2016 US Presidential election and in the Brexit campaign during the same year (Gorodnichenko, Pham, & Talavera, 2018; Howard & Kollanyi, 2016). The purposes for these practices are often the manipulation of media trends and the setting of the pertinent agenda (Bessi & Ferrara, 2016; Bradshaw & Howard, 2017; Ferrara, 2017; Robertson, Riley, & Willis, 2016). These scandals would seem an indication that there are forces acting over Twitter with the ability to weaken user capacities to decide freely about political issues, as previously expressed by Pang and Lee (2008).

However, while the influence of Twitter and other platforms on voting choice has been previously investigated (Correa & Camargo, 2017; Grover, Kar, Dwivedi, & Janssen, 2019; Reed, 2015), it is still unclear how the social media environment, by acting as a stage for factors that influence voting to interact simultaneously and in real time, may actually both confirm and challenge voting intentions of users.

4.3.2. Uses of Twitter as source of data

Academics and campaigners have used Twitter widely to predict user behaviours with the assistance of predictive models (Wicaksono, Vania, Distiawan, & Adriani, 2014). One of the main drivers of engaging on Twitter is that this social media platform offers users the possibility to broadcast tweets to a wider audience inside and outside of their networks. This dissemination feature constitutes a useful source of data that can be more effective for marketing purposes than other platforms (Baltas, Kanavos, & Tsakalidis, 2016; Song & Kim, 2013). In addition, Twitter allows the generation and consumption of vast amount of information, which can be

tracked, recorded, and analysed to build predictive models. Since November 2018⁵, individual tweets have a length limitation of 280 characters. This feature forces users to be as concise as possible (Kurka, Godoy, & Von Zuben, 2017), which helps to reduce vagueness in intended messages. As a result, users' interaction on Twitter can be more heated and critical than on other social media platforms (Wirawanda & Wibowo, 2018), which can also lead to a more polarised environment. In this context, polarisation means that antagonist positions among users are more identifiable.

Twitter can support to increase awareness about a user or topic by reaching a wider audience. To do this, different strategies have been applied, such as engaging influential users in the spreading of content, using Twitter's promote mode⁶ to sponsor tweets, and eye-catching trolling techniques aiming to capture users' attention (Saravanakumar & SuganthaLakshmi, 2012; Thackeray, Neiger, Hanson, & McKenzie, 2008; Younus et al., 2014).

Also, there are cases in which Twitter has supported communication between organisations and users during times of hardship and emergency situations. For example, during critical events such as riots (Panagiotopoulos, Bigdeli, & Sams, 2014; Procter, Vis, & Voss, 2013; Vis, 2013), terror attacks (Eriksson, 2016; Simon et al., 2014), or natural disasters (Ashktorab, Brown, Nandi, & Culotta, 2014), Twitter data have assisted the understanding of user behaviours for both governments/organisations and users, especially when providing resources to deal with the management of these emergencies. Similarly, when negative online content goes viral to affect the reputation of a user or a brand, Twitter has provided campaigners with a platform to respond and implement crisis management instantly, to undo the damage and repair their business names. When no reaction is given to address negative content, as expressed by Mandviwalla and Watson (2014), brands can be severely affected not only in terms of reputation, but also financially.

4.3.3. Twitter in politics

Twitter has been widely used for public and political deliberation about issues of common interest (Auvinen, 2012; Burgess & Bruns, 2012; Hong, Doumith, & Davison, 2013). It has strategically changed how Western and some Eastern politicians run their campaigns and how they interact with other users, making them more accountable and reachable. For example, politicians use Twitter as a platform to

⁵ Originally, the length of characters was set to 140 characters.

⁶ <https://business.Twitter.com/en/solutions/Twitter-promote-mode.html>

engage with users, tailor messages depending on the target audience, weigh public opinion, while also aiming to gain the sympathy of users (Fountaine, 2017). Nowadays, politicians value the importance of establishing their Twitter presence, while carefully considering the image they want to project to others (Gulati & Williams, 2015). Twitter has also increased the possibility of reaching wider audiences, especially when it is perceived that the traditional media are not providing sufficient or fair coverage to candidates (Bitecofer, 2018). It can be also beneficial in relation to how to react during controversial times, or when the prevalent image of the candidates is negative, which can be measured from the public reaction. Therefore, in the task of dynamically shaping the political image of a candidate, information continuously emerging on Twitter may provide sources for effective and opportune action, usually involving adjustments in rhetoric, symbols, and language.

In this sense, one of the first successful electoral campaigns that understood the power of social media platforms, including Twitter, was the one developed by Barack Obama's team during the US Presidential election in 2008. According to Cogburn and Espinoza-Vasquez (2011) and Auvinen (2012), this campaign was successful because of the integration of technologies of Web 2.0, which refers to the applications and content created and distributed in a collaborative approach by users around the world, and allows public deliberation spaces among politicians and voters. Even though Obama was not the first politician to use Twitter and other social media platforms, this campaign highlighted the power of these technologies to disseminate views and attract voters. Similarly, in the 2016 US Presidential elections, Twitter was the main platform Donald Trump used to broadcast his personal and campaign ideas. Beyond the US, politicians in other countries (Auvinen, 2012; Hsu & Park, 2012; Karlsen, 2013; Larsson & Kalsnes, 2014) have also used social media platforms to organise their campaigns while minimising the intervention of third parties such as traditional media or for-profit organisations.

Also, Twitter has functioned as a platform for social mobilisation around the world for protest purposes, even when regimes have tried to control or censor the information that users generate and consume. Political participation has been organised through Twitter by providing and disseminating information in real time, and managing the logistics of the protests (Jost et al., 2018). Prominent examples are the Arab Spring in Tunisia and Egypt (Tufekci & Wilson, 2012), the Occupy Wall Street movement in the US (Juris, 2012), and the *Indignados* protest in Spain (Anduiza, Cristancho, & Sabucedo, 2014). What these mobilisations had in common

is the use of Twitter as a platform to organise and reclaim action, despite a context in which their governments had undermined their complaints and traditional media had given them poor coverage. Thus, the usefulness of Twitter lies in its ability to keep users and outsiders informed and organised about the development of the protests.

Similarly, people engaging in civic and political activities are frequently active users of social media platforms (Pearce & Kendzior, 2012; Tufekci & Wilson, 2012; Valenzuela, 2013; Valenzuela, Arriagada, & Scherman, 2012). It has been assumed that Twitter empowers participatory democratic processes because users are constantly exposed to relevant information (Bavaresco, 2014; Coleman & Gotze, 2001; Lin, Bagrow, & Lazer, 2011; Weiler, 2013). However, this viewpoint has its detractors, who state that social media platforms may weaken democracy because they allow the isolation of users into communities sharing similar viewpoints (Sunstein, 2009). Moreover, it is also believed that the accessibility of public and private information can be used for repressive action (Bradshaw & Howard, 2018). About this, under authoritarian regimes, social media can act as a tool to repress political opponents, implement mass surveillance, and even disseminate extremist information (Morozov, 2011; Pearce & Kendzior, 2012).

4.3.4. What comes after analysis of Twitter data?

When using Twitter, a common aspiration among politicians and campaigners is the ability to produce content that catches the attention of users in a positive way and tempts them to replicate the tweet. This is because positive posts correlate with sales, votes, and share price (Bollen, Mao, & Zeng, 2011; Tuarob & Tucker, 2013; Verma & Sardesai, 2014). Thus, campaigners value the importance of creating and distributing positive eWOM, which can have an influence on electoral outcomes. In this regard, sentiment analysis approaches are often useful to monitor the prevalent feelings of users and adjust the content accordingly.

Despite the popularity of Twitter and its widespread use as a tool for user engagement, little attention has been paid to how it actually influences user behaviours. Some literature on this matter is concerned with the way influential users spread positive eWOM. Yet, the ability to influence has been determined via metrics such as the number of posts and followers for brand awareness, number of replies for engagement, and number of shares for eWOM (Bakshy, Hofman, Mason, & Watts, 2011; Cha, Haddadi, Benevenuto, & Gummadi, 2010; Fan & Gordon, 2014; Jin & Phua, 2014; Kwak, Lee, Park, & Moon, 2010). It is believed that these metrics are a

proxy for the ability to influence behaviour in social media (Vatrapu et al., 2015). However, as stated by Zhang et al. (2015), influence cannot be reduced to these measures, especially when the number of followers or users retweeting content does not necessarily translate into supporting users. Beyond these metrics, it is also important to distinguish among typologies of influential users, understand how they engage with other users, and identify what factors support their popularity.

In addition, tailoring content based on the target audience is another challenge for campaigners when attempting to spread positive eWOM. It means that when producing Twitter content, it is important to match the optimal message to the optimal audience at the optimal time (Penlington, 2017). Previous research has focused on identifying factors that determine the likely volume of retweets. Traditionally, this has involved the use of observable metrics in tweets and the development of predictive models (Jose & Chooralil, 2015; Nesi, Pantaleo, Paoli, & Zaza, 2018; Suh, Hong, Pirolli, & Chi, 2010). However, beyond these metrics, it is likely that predictive models benefit greatly from a rather qualitative analysis of the content to gain deeper insight into the data.

Moreover, modelling users' behaviour is critical when using Twitter as data source. Rost, Barkhuus, Cramer, and Brown (2013) underpinned this issue, and criticised that models of analysis have largely used datasets extracted from social media as a sample representative of the offline world behaviour rather than as communicative and meaning-making data to understand social interactions among users. Similarly, most studies that relied on Twitter data lacked an understanding of the drivers that influenced what was tweeted, which undermines the fact that Twitter data are produced as a side effect of communication between users. Equally, by 2016 Lloyd and Cheshire (2016) had criticised that the value of Twitter as a source of consumer's insights remained under-investigated, although the last two years have witnessed an increasing interest in the subject (Henderson et al., 2017; Rathan, Hulipalled, Venugopal, & Patnaik, 2018; Sun & Rui, 2017). Also, Brooks (2015) has claimed that one of the most relevant challenges for accomplishing a thorough understanding of social media datasets is a present lack of technological support in terms of systems of mixed methods for data analysis (quantitative and qualitative research methodologies), but recent literature has dabbled into mixed methods Twitter research (Amoroso et al., 2018; Komorowski, Do Huu, & Deligiannis, 2018).

Finally, although approaches and applications for sentiment analysis of Twitter content provide an indication of user's satisfaction based on the positivity and

negativity of tweets, it provides little understanding of what users value most. This means that sentiment analysis alone is not a prescriptive tool, for which further analysis is often required to attempt better targeted actions (Moghaddam, 2015). To fill this gap, different predictive models have been applied when a deeper examination is needed (see Table A-4.1 in the Appendix A-4), which often demands analysis and interpretation of Twitter data from different angles (Sarewitz, Pielke, & Byerly, 2000). From the political viewpoint, understanding the reactions of users is critical when designing and implementing public policy, and gaining more support from them.

4.4. Unpacking Twitter analysis

The aim of Twitter analysis is often to measure and boost the impact of content on Twitter. In most cases, this measurement has been based on quantitative metrics such as number of tweets, likes, retweets, or size of networks (Bruns & Stieglitz, 2013; Kruikemeier, 2014). Twitter also offers rich data that is suitable to develop predictive models (Chowdhury, Routh, & Chakrabarti, 2014). Predictive models often support better understanding of Twitter users and allow identification of different factors that can influence their behaviours. Thus, these models are convenient for designing campaigns and making informed decisions. For the development of predictive models, the process usually begins by extracting raw tweets from Twitter, which will later be transformed into a format that is readable for predictive models by using pre-processing approaches. Then, the data are analysed and modelled to develop the predictive models.

4.4.1. Sampling Twitter

From the Twitter context, data sampling refers to the process of selecting a subset of tweets or users to be analysed or used in the development of predictive models (Aghababaei & Makrehchi, 2017). Traditionally, the selection criteria for sampling Twitter have been mainly focused on two perspectives: content-based and user-centric (Aghababaei & Makrehchi, 2017). The former relies on the use of keywords or hashtags to retrieve tweets, while the latter focuses on choosing specific users from the networks, who often feature a high number of followers and are perceived as experts in different fields. Content-based sampling is useful when aiming to extract content from different users with respect to an event of interest (Aghababaei & Makrehchi, 2015), whereas user-centric sampling is useful to discover new and

meaningful patterns from the data (Chepurna, Aghababaei, & Makrehchi, 2015). User-centric sampling is generally perceived by users as a more effective approach for breaking news detection than content-based sampling. This is because users considered as experts are often seen as reliable sources of news (Aghababaei & Makrehchi, 2017). For example, the Twitter user account of a news agency or a news anchor can be seen as a trustful source of information when it comes to breaking news. This claim has been put forward, for instance, in the study done by Zafar et al. (2015) on the 2012 earthquake in Japan.

The decision as to the sampling selection criteria for tweets depends on different factors: for example, the context of the case study, levels of content accessibility based on specific keywords/users, demographic aspects, and target audience. Defining the sampling criteria is an important part of developing predictive models. However, the Twitter literature seems to indicate that this sampling process, rather than being based on guidelines empirically developed, has been mainly conducted intuitively and without a clear rationale. In this sense, academics may find themselves in dilemmas as to which keywords/users are important for sampling purposes. There is also uncertainty on how relevant tweets ignored in the sampling process may affect predictions. Conversely, what happens when tweets that are not relevant to the case study are included simply because they satisfy the keywords? Up to now, far too little attention has been paid to the whole sampling process. A rare exception is the work of Jain and Kumar (2017) on how to extract key information for predictive modelling. Yet, there remains a paucity of evidence on the application of this approach.

Also, little has been said in terms of what constitutes appropriate period or duration of data extraction for predictive modelling purposes. Apparently, this aspect has been addressed, in most cases, intuitively. For example, some academics agree that a 24-hour time window provides the best prediction accuracy when compared to several days of data extraction (Bermingham & Smeaton, 2011; Ramteke, Shah, Godhia, & Shaikh, 2016). However, other studies have relied on longer timespan. Similarly, the way sample size, in terms of number of tweets, influence on the performance of predictive models has not been addressed thoroughly. In this sense, it is necessary to analyse whether addressing the previous questions could boost the power of predictive modelling. More details of the sampling schemes that previous academics have applied to their research are shown in Table A-4.1 and Table A-4.2 in the Appendix A-4.

4.4.2. Pre-processing

The process of transforming unstructured Twitter data into data readable by machine learning approaches is a key challenge when applying Twitter analytics, especially when referring to the use of predictive models (Pandey, Kumar, & Srivastava, 2016; Suthaharan, 2014). Equally challenging is performing the identification of worth extracting data from large volumes (Leskovec, Rajaraman, & Ullman, 2014; Wu, Zhu, Wu, & Ding, 2014), which is sensible when aiming to improve the performance of predictive models to obtain meaningful knowledge useful for campaigners.

Due to the nature of unstructured Twitter data, noisiness and incompleteness bring about a great deal of complexity when aiming to develop predictive models (Hu, Wen, Chua, & Li, 2014; Malik, 2013; Wu et al., 2014) because they can affect the performance of classifiers algorithms (Zhao, 2015; Zhao & Gui, 2017). Noise is defined as the generation of content which is not relevant to the specific case study, and further refers to those tweets that are generated and distributed by fake, bot, and spam accounts (Sprenger, Sandner, Tumasjan, & Welp, 2014). It is believed that noisiness increases the dimensionality of text, reducing in this way the overall performance of predictive models while also delaying the speed-up of the process (Haddi, Liu, & Shi, 2013). On the other hand, incompleteness refers to the presence of poorly structured content (Zhao, Gui, & Zhang, 2018), which in the case of Twitter takes on more relevance due to the limitation of 280 characters. Examples of incompleteness are the use of abbreviations, grammatical variance, slang, or “mundane chatter” (Balahur, 2013; Burnap & Williams, 2015), which challenges the modelling process.

To reduce the noisiness and dimensionality of the data, pre-processing approaches have been applied to clean the datasets prior to the modelling process (Wu et al., 2014). Pre-processing involves several sub-steps such as the removal of extra white spaces, numbers, stop-words, or transformations of other words. It also involves extracting feature vectors aiming to make the data readable before the implementation of models. Several techniques have emerged for this purpose, such as the application of tokens, n-grams, part-of-speech tagging (POS), or the stemming process. Tokenisation is a segmentation process where the text is split into individual words by white spaces, line breaks, or punctuation signs, making the text into a bag-of-words (BOW) that allows the feature extraction. N-gram refers to the identification of

sequences of tokens, such as unigram (1 token), bigram (sequence of two tokens), or tri-gram (sequence of three tokens). POS allows the identification of the structural elements of text, such as noun, verb, and adverb. And, stemming reduces the variants of a word into its stem by getting rid of prefixes and suffixes. The construction of feature extraction influences the overall performance of a classifier algorithm (Agarwal et al., 2011; Bhadane, Dalal, & Doshi, 2015; Pak & Paroubek, 2010; Ruba & Venkatesan, 2015; Smailović, Grčar, Lavrač, & Žnidaršič, 2013), which can be designed and personalised depending on the data and requirements. The implementation of the pre-processing phase is relevant because it can enhance the accuracy of the classifier.

When pre-processing tweets, there are some features of tweets that have been subject to transformation processes. For example, replacing emoticons with their equivalent words has been applied to build some predictive models. As shown in Table A-4.1 and Table A-4.2, other features such as hashtags, URLs, and mentions have been usually discarded from datasets or undermined from analysis, claiming that these contribute to noisiness and overfitting (Al Hamoud, Alwehaibi, Roy, & Bikdash, 2018; Jain & Jain, 2019; Pak & Paroubek, 2010; Pandarachalil, Sendhilkumar, & Mahalakshmi, 2015).

However, these features are very popular among Twitter users because they can provide advantages in terms of more visibility, increase users' interaction and attention, generate positive eWOM, and draw more followers (Ince, Rojas, & Davis, 2017). Hence, the claim that these features do not add value for predictive analytics does not seem to be factually based. Instead, it is likely that these features influence the strength of the predictive models. In this sense, there are still opportunities to continue developing and improving classification modelling, either by focusing on qualitative analysis, which could allow campaigners to understand in a deeper way what is influencing user behaviour based on the information they consume on Twitter, or by strengthening the existing quantitative models by including new variables depending on the context.

4.4.3. What analytical tools are recommended to meet the aims?

Twitter-based predictive analytics have surged as key platforms for supporting the understanding of user interests and for predicting user behaviours across a wide range of activities in human and mechanical systems (Bollen et al., 2011; Burnap & Williams, 2015; Hong et al., 2013). In this sense, the application of classification-

predictive models is one of the most utilised analytical tools for academics and campaigners. Classification refers to the process of predicting the class of an observation based on a set of connected variables. This approach has reported accuracy in determining sentiment polarization and classifying political membership and party affiliation of users (Boutet & Yoneki, 2011; Pennacchiotti & Popescu, 2011a; Preoțiuc-Pietro, Liu, Hopkins, & Ungar, 2017), as well as cyber hate speech (Burnap & Williams, 2015), spam detection (Chen et al., 2015a; Wang, 2010), and crime incidents (Chen, Cho, & Jang, 2015b; Wang, Gerber, & Brown, 2012). Sarewitz et al. (2000) explain that predictive models have traditionally served two purposes: first, to test scientific understanding or hypotheses that can help in explaining events or how things work in real life; and second, to serve as a foundation to guide the decision processes in different fields.

In reference to the former purpose of predictive models expressed by Sarewitz et al. (2000), Twitter has been widely used to detect sentiment in a wide range of situations to overall determine users' feelings towards products, services, or users. Twitter provides a rich source of data for sentiment analysis because of the amount of information generated and consumed by users from different backgrounds (Pak & Paroubek, 2010). Sentiment analysis, also known as opinion mining, is described as the process to identify and classify text into different polarity classes depending on the detected feeling and can combine the use of natural language processing (NLP) with machine learning approaches to determine the prevalent sentiment. This classification can be performed using either binary, such as "positive" and "negative" feelings, or multi-class sentiments including "neutral" or more extreme feeling rating levels, for example "very positive", "very negative", or "none" classifications. As for the latter purpose, modelling Twitter data can support different types of decisions, for example, adapting business strategy after discovering and understanding users' preferences and supporting marketing campaigns (Culnan et al., 2010; Hong et al., 2013; Mihalcea & Savulescu, 2013).

Building classification-predictive models using Twitter starts by retrieving tweets. Data can be obtained by downloading pre-existing datasets⁷ or by creating new ones using keywords as explained in Subsection 4.4.1. If one opts for the latter option, data extraction can be performed by using the Twitter Streaming⁸ and Search

⁷ https://dataturks.com/projects/Trending?type=TEXT_CLASSIFICATION

⁸ <https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data.html>

API⁹ (application programming interface), in conjunction with statistical software such as R Core Team (2013) or Python (2019), which can also be used for the modelling phase. Then, the data pre-processing stage is applied to transform data into a readable format for modelling purposes. Usually, this leads to an explanatory predictive model used to forecast outcomes of the phenomenon of interest based on empirical data, making it possible also to detect new patterns or behaviours which can be applied for supporting theory building, theory testing, and relevance assessment (Lassen, la Cour, & Vatrapu, 2017; Shmueli & Koppius, 2011). Table A-4.2 in the Appendix A-4 presents a summary of previous works that have relied on Twitter analysis, showing how the Twitter analysis was conducted and the outcomes of these studies.

A critical component for developing sentiment analysis models is the availability of labelled corpora to classify sentiment. There are different pre-established English language corpora available for application in sentiment analysis such as the *Bing Liu Opinion Lexicon* (Hu & Liu, 2004), *SentiWordNet* (Baccianella, Esuli, & Sebastiani, 2010), *Vader Lexicon* (Hutto & Gilbert, 2014), emoticon-based (Go, Huang, & Bhayani, 2009; Pak & Paroubek, 2010), or the German-based *SentiWS Lexicon* (Remus, Quasthoff, & Heyer, 2010), *SentiML* for Italian language (Di Bari, Sharoff, & Thomas, 2015), *SALDO Lexicon* for Swedish (Nusko, Tahmasebi, & Mogren, 2016), and an Indonesian corpus developed to overcome limitations for under-resourced languages (Wicaksono et al., 2014), to name but a few. Some of these corpora are also available for implementation using statistical computing programming languages. The development of a corpus for sentiment purposes can be carried out by experienced annotators, who need to manually label the detected feeling, or by using a labelled seed corpus (Wicaksono et al., 2014) with a small collection of positive and negative tweets, that will expand when building the model.

However, a major limitation arises when there is limited availability of robust corpora to develop the sentiment classifier, especially when working in non-English languages. The development of a labelled corpus to support tweet polarity classification can be time-consuming and expensive due to the voluminous content found on Twitter (Wicaksono et al., 2014), while also requiring extensive manual annotation from humans. In the absence of corpora, some studies have translated text from other languages into English to be able to use pre-existing dictionaries to build a

⁹ <https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html>

classifier (Abdalla & Hirst, 2017; Agarwal et al., 2011; Araujo, Reis, Pereira, & Benevenuto, 2016; Balahur, Mihalcea, & Montoyo, 2014; Joshi, Balamurali, & Bhattacharyya, 2010). However, translation quality might be affected during this process, altering the overall sentiment of text, while sparseness and noise are added to data (Balahur & Turchi, 2012; Lohar, Afli, & Way, 2017; Pennebaker, Booth, & Francis, 2007).

In the absence of robust corpora to conduct sentiment analysis, an alternative approach is the use of text analytics and sentiment analysis providers. This alternative is also valuable when users are not familiar with pertinent programming skills. These providers allow end-users to proceed with sentiment analysis through their APIs or add-in tools, while also letting their customers personalise dictionaries to adapt to local contexts. Examples of commercial providers include *Semantria*¹⁰ supporting 25 languages for sentiment analysis, *MeaningCloud*^{TM11} with 6 languages available, or *SentiStrength*¹² (Thelwall et al., 2010) with 15 languages. Open source tools are also available, such as *GATE* developed by the University of Sheffield¹³ in the UK supporting 12 languages, or the *Stanford Sentiment Analysis Module*¹⁴ including 7 languages available for sentiment purposes. Both commercial and open source providers have been widely used to support political studies, as detailed in Table A-4.2. In addition, a broad summary of election predictions using sentiment analysis can be found in the study conducted by Jain and Kumar (2017).

4.5. Challenges and critics of the use of Twitter data

This paper has so far reviewed the use of Twitter as data source for developing predictive models to understand and analyse user interests and behaviours. Now, the discussion shifts to the limitations associated to Twitter data for the same purpose.

Perhaps one of the most relevant criticism concerns the ethics of handling Twitter data for research purpose. From the perspective of Webb et al. (2017) and Williams, Burnap, and Sloan (2017), ethical issues arise when sensitive tweets are quoted and used in publications without proper consent of their authors, as in the case where Innes, Roberts, Preece, and Rogers (2018) were criticised after publishing their study. Even though they published the tweet content without Twitter handles

¹⁰ <https://www.lexalytics.com/technology/sentiment-analysis>

¹¹ <https://www.meaningcloud.com/developer/sentiment-analysis>

¹² <http://sentistrength.wlv.ac.uk/>

¹³ <https://gate.ac.uk/>

¹⁴ <https://nlp.stanford.edu/sentiment/>

“@username” the content could be still traceable, and *ipso facto* so could its authors. However, in the view of Zimmer and Proferes (2014) these issues do not constitute an ethical problem since it is stated that public tweets are available to third parties, including for academic and commercial purposes. The dissemination of tweets can be made only when Twitter user accounts are set up as public, which is the default setting when opening a Twitter user account. Some restrictions are specified in the Twitter User Agreement¹⁵, but these aside, all public Twitter content can be accessed and distributed by other Twitter users.

Concerning user privacy-related considerations, with the aim of protecting European Union (EU) and European Economic Area (EEA) users’ rights, and to provide a regulation for international business regarding transparency of the use of online data, in May 2018 the General Data Protection Regulation (GDPR) was implemented. This means that, supposedly, users are aware of how their personal data are used, processed, analysed, and for what purposes, even outside the EU and EEA areas. Even though there are complaints about violations of the GDPR, especially when social media platforms refuse to give users information about how their data are used (Meyer, 2018; Novak, 2018). The need for this regulation is explained in that companies have misused data in the past without users’ consent. For example, Cambridge Analytica, a British political consulting firm, used data from Facebook to target users with personalised political advertisements during political campaigns, without users’ authorisation of their data being used for political purposes (Persily, 2017; Punit, 2018). However, for the case of social media platforms, consent relating to data extraction and analysis is given when users accept the terms and conditions of use, as is the case for Twitter¹⁶:

By submitting, posting or displaying Content on or through the Services, you grant us a worldwide, non-exclusive, royalty-free license (with the right to sublicense) to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute such Content in any and all media or distribution methods (now known or later developed). This license authorizes us to make your Content available to the rest of the world and to let others do the same. You agree that this license includes the right for Twitter to provide, promote, and improve the Services and to make Content submitted to or through the

¹⁵ <https://help.Twitter.com/en/rules-and-policies/Twitter-rules>

¹⁶ <https://Twitter.com/tos?lang=en#us>

Services available to other companies, organizations or individuals for the syndication, broadcast, distribution, promotion or publication of such Content on other media and services, subject to our terms and conditions for such Content use. Such additional uses by Twitter, or other companies, organizations or individuals, may be made with no compensation paid to you with respect to the Content that you submit, post, transmit or otherwise make available through the Services.

From the methodological viewpoint, criticism arises from the restriction of the Twitter Search API when extracting tweets to an external source, since it is set to allow only the 3200 most recent tweets every time. This might limit the sampling volume. Data sampling using the standard package, which does not have any cost, is focused on relevance and not on completeness¹⁷, meaning that, based on a specific query for data extraction, not all matched tweets will be retrieved. In addition, there is another limitation in that the Twitter Search API allows users to retrieve tweets produced no more than one week before. However, if completeness is needed, Twitter provides the premium and enterprise API services at a cost, depending on the user's needs. Both services make it possible to either retrieve tweets posted within the last 30 days, or the full-archive collecting tweets from 2006, and the difference lies in the level of access and reliability needed, which is superior in the enterprise package. However, the evaluation of random sampling through the Twitter Search API compared with the Twitter Stream Firehose, which retrieves all the tweets meeting the selection criteria, concludes that even when the coverage of the former is smaller in random samplings, if using specific parameters this approach can provide as valuable data as the Firehose method (Morstatter, Pfeffer, Liu, & Carley, 2013), overcoming to some extent the fear of limits to sampling size.

Another challenge in using Twitter data as a source derives from the level of penetration, meaning that in some countries the accessibility and popularity of Twitter might not be relevant for academic or political purposes, especially in non-democratic societies or in countries where Western social media platforms are not frequently used (Reuter & Szakonyi, 2015). For such cases, the uses of other social media platforms where public issues are discussed is an alternative for data sampling. In addition, the use of traditional means of data sampling is critical when sensitive topics need to be

¹⁷ <https://developer.Twitter.com/en/docs/tweets/search/overview/standard.html>

covered such as income, sexual orientation, religious beliefs, or catastrophic diseases that might not be openly discussed through Twitter. In addition, some problems might arise when there is a lack of infrastructure and human capacity to implement Twitter solutions, in which circumstances outsourcing these functions to third-party providers can be a solution.

Also, criticism relates to the presence of fake news and accounts, such as bots and trolls, which aim to mislead users. In this sense, previous works have focused on the use of automatic detection methods for finding fake news and identifying fake accounts (Conroy, Rubin, & Chen, 2015; Rubin, Conroy, Chen, & Cornwell, 2016; Shu et al., 2017; Spohr, 2017; Wang, 2017). It is believed that they have the power to impact on the economy in the offline world (Cresci et al., 2015; Gupta, Lamba, & Kumaraguru, 2013), while also influencing users' behaviour. For example, in politics it is claimed that during the 2016 US Presidential election the Russian government used trolls to disseminate fake information to mislead social media users¹⁸. Similar claims of interventions in other countries' elections have recently come to light¹⁹. However, there is still a lack of evidence sufficient to assert that this interference effectively influenced the decision process, and to what extent these can constitute a threat to democracies. In this regard, future research should look at how these entities can affect user behaviour-predictive models.

In addition, representativeness has been discussed and debated by academics as a limitation, explaining that Twitter is neither representative of the offline population nor of Twitter users (Asher, Leston-Bandeira, & Spaiser, 2019; Blank, 2017; Karusala, Kumar, & Arriaga, 2019; Mellon & Prosser, 2017). This argument is based on the limitation of social media demographic representativeness, meaning that there might be the exclusion of certain users who do not use Twitter for public discussions or disadvantage to minorities. For example, studies have concluded that politically active Twitter users tend to be male, younger, wealthier, more educated than the general population, and located in inner-city areas (Anderson, 2018; Malik, Lamba, Nakos, & Pfeffer, 2015; Mellon & Prosser, 2017). Twitter representativeness reflects that data might be biased due to population characteristics. However, this should not be always seen as a limitation, since in some fields, such as in marketing, it is desirable to target a specific audience based on age, socio-economic, or education level.

¹⁸ <https://www.nytimes.com/news-event/russian-election-hacking>

¹⁹ <https://www.business-humanrights.org/en/archimedes-group>

Finally, as previously referred to in Subsection 4.3.1, another criticism associated with the use of Twitter relates to social network homophily (Asher et al., 2019; Halberstam & Knight, 2016; Karimi et al., 2018; McPherson et al., 2001). Homophily, also called cyber-balkanisation, is present in the seeking out of information to reinforce users' previous beliefs, disregarding those arguments that challenge or influence their current viewpoints, even in the presence of robust evidence (Chan & Fu, 2017). In this sense, when users have positions, especially on sensitive topics, it is believed that the role Twitter plays is to polarise and reinforce existing beliefs such that they become more extreme (Halberstam & Knight, 2016). On the other hand, when users do not clearly have standpoints, Twitter might have a greater influence on the decision process. However, more information on the influence of Twitter on polarising opinions would help research to establish a greater degree of accuracy in this matter.

4.6. Conclusion

This review has identified use and limitations of Twitter when used as data source for modelling purposes. Concerning use, Twitter data have been used widely to develop both sentiment analysis and predictive models. Twitter-based sentiment analysis has been useful to identify users' opinions, and in the political context, also to measure vote share of candidates during elections. Although Twitter-based sentiment analysis has not played a prescriptive role for analysts, it has provided understanding about user preferences and behaviours.

As for the development of predictive models, the emergence of Twitter has brought new light to the way user behaviours can be analysed. Twitter lays bare the presence of under-explored factors that might have an impact on users' decision processes. In this regard, although it remains undetermined, the literature suggests that different types of choices are being influenced by the information that users consume on social media every day. Thus, identifying and understanding such factors can provide new resources for better targeted marketing strategies and user relationship management. In this sense, Twitter is a powerful research tool because it provides academics and campaigners with a platform to extract and obtain users' insights to be transformed into knowledge.

Finally, regarding limitations, criticism has been based on how Twitter data compare to data obtained by traditional means, such as surveys or questionnaires. Particular attention has been paid to the weakness of Twitter to investigate sensitive

topics that rely on confidential data. In this regard, traditional means of data sampling allow more privacy control of user information when compared to Twitter, and thereafter, are commonly used by academics. Likewise, Twitter is less effective as data source in societies where the level of penetration of Twitter and other social media platforms is low. Despite these limitations, the literature seems to suggest that Twitter is convenient when it comes to easiness of data sampling.

4.7. Further research path

The relevance of this chapter is the identification of areas for future research in the use of Twitter-based data for modelling purposes. Even though the study of predictive models using Twitter has been covered in previous years, there are several questions that still remain to be answered. First, the issue of sampling. Previous studies have developed tweet selection criteria for data sampling in an intuitive way. This means that there are no specific criteria when choosing keywords or users to gather data. This might constitute a limitation when building predictive models because relevant information can be either discarded from analysis or included increasing the noisiness of the data. Similarly, there are no guidelines concerning the number of tweets retrieved and the timeline for data extraction. Concerning the timeline, there are studies that have relied on 24-hour data extraction, while others have used data retrieved from archives going back years. So, the development of criteria for tweet selection seems to be a relevant issue that needs much more attention.

A second issue relates to data pre-processing. This study has presented a review of previous research showing that most past work has discarded some features that are embedded in tweets, such as hashtags or URLs, claiming that they do not add value to the performance of models but instead contribute to the noisiness and dimensionality of datasets. Sometimes, these features have been analysed from a quantitative perspective by counting their presence in tweets. However, these features can contain information that can provide better insights into users' behaviours, allowing the discovery of new patterns and supporting communication strategies based on the target audience. Hence, further research should extend the scope of data pre-processing usefulness on Twitter sentiment classification by considering the analysis of feature selection of tweets in the models. Similarly, pre-processing should also address data quality issues in unstructured data such as noisiness and incompleteness, as reviewed in Subsection 4.4.2.

A third issue concerns analytical tools that support the development of predictive models. Previous research has mainly used sentiment analysis tools to determine users' feelings towards entities, and mainly in the English language. In this sense, it is important to continue improving predictive modelling and developing robust corpora to build sentiment analysis in non-English languages.

Fourth, the use and impact of influential users on outcomes have not been investigated in depth. The literature has focused on defining and identifying who they are by considering metrics such as the number of followers or the power to spread content among networks using retweets. These metrics are not exhaustive, but they provide campaigners with a useful tool to analyse the impact of targeted marketing initiatives. However, future work can be focused on developing influential users' typologies to support the understanding of what makes these users influential beyond the metrics previously exposed. This can be oriented to designing strategies as to the criteria of whom to hire and work with, aiming to maximise offline outcomes. The development of classification schemes covering these users will depend on the desired target. Therefore, this typology provides campaigners with a tool to increase awareness and reach strategic target users based on their needs, while supporting strategies aiming to build stronger connections with other users.

Finally, and not less important, there are the effects of fake users and misleading content. This is especially pressing in political matters. In fact, it is still a point of debate to what extent they can actually change or polarise users' voting behaviours. Since Twitter is reshaping political landscapes, it is relevant to continue examining how the exposure of fake news and accounts can influence these aspects of decision behaviour, while determining to what extent they can become a threat to democratic societies. Similarly, it is important to conduct future research into the influence of homophily and confirmation bias to determine if these aspects can change voting behaviour or make polarisation more extreme.

Appendix A-4.

Table A-4.1

Summary of studies using Twitter data for classification-predictive modelling

Authors	Aims of using Twitter	Outcome	Sampling criteria	Pre-processing and feature extraction	Model implementation	Take away for supporting campaigners
Bakshy et al. (2011)	Identification of influential users on Twitter.	Influential users tended to have larger number of followers, while URLs with a positive connotation were more shared by users.	Selection criteria: Tweets including URLs about specific events. Timeline: 2 months. Tweets extracted: 74 million.	Not specified.	Manual classification of URL content was implemented by using the crowdsourcing marketplace <i>Amazon's Mechanical Turk</i> ²⁰ . Regression tree model was applied for modelling purposes.	Campaigners could build relationships with influential users to spread content, reach new audiences, generate business, and create partnerships with them.
Pennacchiotti and Popescu (2011b)	Classification of users depending on their demographics, political and marketing interests.	Identification and detection of political affiliation, ethnicity, and affinity to a specific business was possible using Twitter content.	Selection criteria: Tweets containing specific keywords previously defined by the study. Timeline: Not specified. Tweets extracted: Not specified.	Not specified.	Entity classification was implemented by using <i>OpinionFinder</i> ²¹ , while gradient-boosted decision tree was used to implement the model.	Results from this study could be used to implement and tailor marketing and political campaigns depending on the target audience.
Ribeiro et al. (2012)	Modelling traffic conditions and incidents.	The study found a significant correlation between the real traffic conditions and tweets.	Selection criteria: Tweets containing mentions of user accounts responsible for informing about	Accent marks, URLs, and mentions were removed. Retweets were excluded from the datasets.	Some tweets were manually labelled according to the geographic sector, and traffic conditions. Then, exact and approximate	This research provides critical information to design and support the logistics of traffic management.

²⁰ <https://www.mturk.com/>

²¹ <https://mpqa.cs.pitt.edu/opinionfinder/>

Authors	Aims of using Twitter	Outcome	Sampling criteria	Pre-processing and feature extraction	Model implementation	Take away for supporting campaigners
			traffic conditions. Timeline: 3 months. Tweets extracted: 10,005.		string-matching approaches were used for modelling purposes.	
Adrien, Cécile, and Hakim (2013)	Prediction of information propagation.	The proposed model could predict the dynamics of the diffusion of information but failed when forecasting the volume of tweets.	Selection criteria: Tweets gathered from Yang and Leskovec (2011), and topology of the network from Kwak et al. (2010). Timeline: 7 months. Tweets extracted: 467 million.	For each user, followers distant was calculated and connected with their following relationships to construct the cascades.	Decision tree, linear and multilayer perceptron, and Bayesian Logistic regressions were used to implement the model.	This study could provide campaigners with a tool to develop campaign strategies based on the identification of features that influence the propagation of information.
Hong et al. (2013)	Modelling users' behaviour.	Twitter data were useful when modelling users' behaviour, especially oriented to predicting whether a tweet would be retweeted by a target user.	Selection criteria: Tweets from users who had posted at least 10 tweets and including one retweet. Timeline: 1 month. Tweets extracted: Approximately 11 million.	Features related to categorical features, content profiles, relevance scores, latent topic model, and content meta-features were used for analysis and modelling purposes.	Co-factorisation machines were used to implement the model.	Understanding and modelling users' behaviour and their connections is critical to revealing how users interact with others.
Adrover et al. (2015)	Identification of adverse effects of HIV drug treatment.	Twitter accurately represented adverse effects associated with some of the drugs used to treat this disease.	Selection criteria: Tweets were purchased from Gnip Inc ²² . Criterion was to select tweets that included antiretroviral	Tokenisation was used to pre-process the data. Tweets labelled as noisy were discarded for later analysis. Tweets were quantitatively	Machine learning methods such as decision trees, SVM (support vector machine), and artificial neural networks were implemented for modelling purposes. Sentiment analysis was manually	The results from this study could support the design of strategies to understand the role drug treatment plays for epidemiological

²² <http://support.gnip.com/>

Authors	Aims of using Twitter	Outcome	Sampling criteria	Pre-processing and feature extraction	Model implementation	Take away for supporting campaigners
			keywords. Retweets were not considered. Timeline: 3 years. Tweets extracted: Approximately 40 million.	measured by the number of features available.	performed.	diseases.
Burnap and Williams (2015)	Detection of cyber hate speech.	Diffusion of cyber hate on Twitter might trigger hate crimes in the offline world.	Selection criteria: Tweets containing keywords about a specific criminal event. Timeline: 2 weeks. Tweets extracted: Approximately 450,000.	BOW (bag-of-words), tokenisation, stemming, and n-grams techniques were implemented. Words were transformed to lower case. Non-alphanumeric characters, emoticons, and punctuation signs were removed.	Tweets were partially manual labelled to identify online hate speech using <i>CrowdFlower</i> , now called <i>Figure Eight</i> ²³ , as the crowdsourcing resource. Use of <i>Stanford Lexical Parser</i> (De Marneffe, MacCartney, & Manning, 2006) to extract dependencies. Decision tree, and SVM approaches were used for classification purposes.	The model could provide campaigners with a tool for detecting cyber hate, so early responses can be taken to reduce criminal activities.
Eichstaedt et al. (2015)	Prediction of heart disease mortality.	Language used on Twitter revealed psychological characteristics that were related with heart disease mortality rate.	Selection criteria: Tweets were obtained from a sample provided by Twitter called the “Garden Hose” ²⁴ . Timeline: 10 months. Tweets extracted: 826,000.	Use of automatic process to extract frequency of words and phrases was implemented before the modelling phase.	LIWC2007 (Pennebaker, Boyd, Jordan, & Blackburn, 2015) was used for emotion classification, and linear regression to fit the model.	Findings are vital to predict heart disease vulnerability among Twitter users.

²³ <https://www.figure-eight.com/>

²⁴ <http://www.sobigdata.eu/content/twitter-stream-gardenhose-daily-access>

Authors	Aims of using Twitter	Outcome	Sampling criteria	Pre-processing and feature extraction	Model implementation	Take away for supporting campaigners
Hoang and Mothe (2017)	Prediction information diffusion through retweets.	Number of followers, followees, and the number of groups that the user belongs to, were some features that allowed the prediction of whether a tweet would be retweeted, and if so, the volume of retweets to be received.	Selection criteria: Tweets were obtained from Tamine et al. (2016). Timeline: 3 days. Tweets extracted: 500.	User-based, time-based, and content-based features were extracted and Boolean or numerically codified.	Sentiment analysis was applied using the corpus from <i>SemEval-2013</i> ²⁵ , and from sentiment movie reviews ²⁶ . Machine learning approaches such as Naïve Bayes, SVM, and Random Forest were implemented for modelling purposes.	The study could be helpful to understand and control the diffusion of information on Twitter.
Alp and Ögüdücü (2018)	Prediction of influential Twitter users.	Twitter made it possible to identify and model information diffusion within networks.	Selection criteria: Tweets and network information generated by specific users. Timeline: 69 days. Tweets extracted: Not specified.	Stemming was applied to the dataset. Then, stop-words, punctuation signs, and mentions were removed.	Latent Dirichlet Allocation was applied for topic modelling purposes. For user modelling, user features that were correlated with being influential were identified through matrix factorisation.	Identification of influential users could allow campaigners to develop strategies either to maximise the spread of information through them or to minimise their influence.

²⁵ <https://www.cs.york.ac.uk/semeval-2013/>

²⁶ <https://pythonprogramming.net/new-data-set-training-nltk-tutorial/>

Table A-4.2

Summary of studies that use sentiment analysis of Twitter data about politics

Authors	Aim of using Twitter	Outcome	Sampling criteria	Pre-processing and feature extraction	Model implementation	Model performance evaluation	Language
Tumasjan, Sprenger, Sandner, and Welpe (2010)	Prediction of the 2009 German Federal election.	The number of tweets reflected the election results.	Selection criteria: Tweets including the names of the different political parties involved in the election. Timeline: 37 days. Tweets extracted: 104,003.	Pre-processing and feature extraction are not specified.	<i>LIWC2007</i> software (Pennebaker et al., 2007) was used to determine polarity of tweets.	Twitter results provided a higher accuracy than traditional election polls.	German tweets were translated into English.
Choy, Cheong, Laik, and Shung (2011)	Prediction of the 2011 Singaporean Presidential election.	Tweets could not predict the actual winner of the election.	Selection criteria: Tweets containing candidates' names. Timeline: 8 days. Tweets extracted: 16,616.	Pre-processing and feature extraction are not specified.	A customised corpus was developed to determine feelings of tweets.	Not specified.	English.
Sang and Bos (2012)	Prediction of the 2011 Dutch Senate election.	The number of tweets predicted some seats for the election.	Selection criteria: Tweets including the names of the political parties involved. Timeline: 6 days. Tweets extracted: 64,395.	Pre-processing and feature extraction are not specified.	Dutch political corpus was manually built for sentiment analysis purposes.	Sentiment analysis improved prediction. Count of tweets was closer to election polls results.	Dutch.
Bakliwal et al. (2013)	Development of a sentiment classifier based on the 2011 Irish General election.	Twitter provided a strong dataset for political sentiment classifier generation.	Selection criteria: Tweets containing keywords about the main political entities. Timeline: 5 days. Tweets extracted:	POS tagging from Gimpel et al. (2010), and parsing from Klein and Manning (2003). n-gram was	Manual labelling to determine sentiment of tweets. <i>Subjectivity Lexicon</i> (Wilson, Wiebe, & Hoffmann, 2005), and <i>SentiWordNet</i> (Baccianella et al., 2010) were	Twitter features in the study using supervised machine learning approach provided a better performance than unsupervised approaches.	English.

Authors	Aim of using Twitter	Outcome	Sampling criteria	Pre-processing and feature extraction	Model implementation	Model performance evaluation	Language
			2,624.	implemented.	implemented, while SVMLight (Joachims, 1999) was used to build the model.		
Makazhanov, Rafiei, and Waqar (2014)	Prediction of users' political preferences during the 2012 Alberta, and the 2013 Pakistani, General elections.	Twitter content and users' behaviour could be used for predicting political preferences of users.	Selection criteria: Tweets including keywords concerning candidates' names and elections. Non-personal accounts, such as from the media, business, or fan clubs were removed from the datasets. Timeline: 10 days. Tweets extracted: 181,972.	Pre-processing and feature extraction are not specified.	<i>SentiStrength</i> (Thelwall et al., 2010) and Naïve Bayes were used for classifying polarisation of tweets. Decision Tree and Logistic regression approaches were implemented for modelling purposes.	Model outperformed in comparison with other baselines, including those from human annotators.	English.
Dwi Prasetyo and Hauff (2015)	Prediction of the 2014 Indonesian Presidential election.	Twitter was a reliable predictor of the presidential election.	Selection criteria: Tweets originated in Indonesia containing keywords about candidates' names. Twitter user accounts manually labelled as non-human were discarded. Timeline: 84 days. Tweets extracted: 7,020,228.	Mentions, URLs, electoral keywords, emoticons, and single characters were removed.	Emoticons were used to build the corpus for the sentiment analysis. Naïve Bayes was used for implementation.	Twitter data provided a more accurate result compared to most traditional offline polls.	Indonesian.
Jose and Chooralil	Measurement of political	Twitter provided accurate	Selection criteria: Tweets including	Removing hashtags, URLs,	<i>SentiWordNet</i> (Baccianella et al., 2010) was applied to	Accuracy of Twitter prediction, with 78.6%,	English.

Authors	Aim of using Twitter	Outcome	Sampling criteria	Pre-processing and feature extraction	Model implementation	Model performance evaluation	Language
(2015)	sentiment from the 2015 Delhi Legislative Assembly election.	predictions for taking the most seats in the election.	keywords about the two presidential candidates. Timeline: 3 weeks. Tweets extracted: 12,000.	mentions, and special characters. Tokenisation and speech tagging were performed.	determine polarity of tweets.	was higher than the model proposed by Khan, Bashir, and Qamar (2014).	
Burnap et al. (2016)	Forecast the 2015 UK General Election.	Tweets could not forecast the party holding most Parliamentary seats after the election.	Selection criteria: Tweets including the name of the parties or leaders' names. Timeline: 101 days. Tweets extracted: 13,899,073.	Pre-processing and feature extraction are not specified.	Sentiment analysis was implemented using software from Thelwall et al. (2010).	Results from the election showed that the winning party was the Conservatives, instead of Labour which was the Twitter forecasting result.	Not specified.
Ramteke et al. (2016)	Prediction of the 2016 US Presidential election.	Twitter predicted the winner of the election.	Selection criteria: Tweets including the name of the two candidates and their parties. Timeline: 1 day. Tweets extracted: 121,594.	Mentions and URLs were removed from tweets. Treatment of hashtags and emoticons are not detailed.	TF-IDF (term frequency - inverse document frequency) was applied to identify relevant terms associated with sentiment. Manual labelling using hashtag clustering and automated classification tool through VADER (Hutto & Gilbert, 2014) were implemented. Naïve Bayes and SVM were applied for model implementation.	Twitter results were consistent with the election.	Not specified.
Sharma and Moh (2016)	Prediction of the 2016 Indian election.	Twitter data could predict the winner of the election.	Selection criteria: Tweets including hashtags related to five Indian political parties. Timeline: 1 month.	URLs, hashtags, mentions, stop-words, emoticons, and special characters were removed	TF-IDF transformation was applied. Unsupervised learning using dictionary based on the use of <i>SentiWordNet</i> for Indian languages (Das &	Twitter results were consistent with the election.	Hindi.

Authors	Aim of using Twitter	Outcome	Sampling criteria	Pre-processing and feature extraction	Model implementation	Model performance evaluation	Language
			Tweets extracted: 42,235.	from the dataset. n-gram was implemented.	Bandyopadhyay, 2010), and manual labelling was partially performed. Also, supervised learning through Naïve Bayes and SVM were applied for sentiment implementation.		
Tunggawan and Soelistio (2016)	Prediction of the 2016 Primary election in the US.	Twitter data did not predict the winner of primary elections, neither for the Republicans nor for the Democrats.	Selection criteria: Tweets including a hashtag related to the election. They were then filtered if they contained the name of any of the candidates. Timeline: 75 days. Tweets extracted: 371,264.	URLs were removed for the analysis. Hashtags, mentions, and retweets were not processed from the dataset.	Tweets were manually labelled to identify the candidate and sentiment. Naïve Bayes was used for modelling prediction purposes.	Twitter forecast could not give accurate predictions of the election, whereas professional poll organisations did.	English.
Anuta, Churchin, and Luo (2017)	Prediction of the 2016 US Presidential election.	Twitter data could predict the winner of the election.	Selection criteria: Secondary tweets from Littman, Wrubel, and Kerchner (2016). Timeline: 4 months. Tweets extracted: Approximately 3 million.	Not specified.	Lexicon VADER (Hutto & Gilbert, 2014) was used for sentiment analysis purposes.	Twitter provided a worse predictive performance than traditional polls.	Not specified.
Jain and Kumar (2017)	Prediction of the 2015 Delhi Assembly election.	Twitter predicted the winning party with most seats for the election.	Selection criteria: Tweets containing keywords about the election. Timeline: 79 days. Tweets extracted:	Tokenisation, stop-word removal, stemming, and dimensionality reduction, among	Several supervised learning approaches were used for implementation, such as SVM, decision tree, and Naïve Bayes.	Twitter data outperformed in comparison with some traditional polls.	Not specified.

Authors	Aim of using Twitter	Outcome	Sampling criteria	Pre-processing and feature extraction	Model implementation	Model performance evaluation	Language
			703,521.	others, were applied before the sentiment analysis.			
Budiharto and Meiliana (2018)	Prediction of the 2019 Indonesian Presidential election.	The study predicted the winner of the election.	Selection criteria: Tweets including hashtags about the election. Timeline: 5 months. Tweets extracted: Not specified.	URLs, stop-words, and special characters were removed from the datasets.	Manual training was conducted on a small subset of data, while also using <i>TextBlob</i> (Loria et al., 2014) for polarity classification.	Twitter results were consistent with traditional polls.	Indonesian.

References

- Abdalla, M., & Hirst, G. (2017). Cross-lingual sentiment analysis without (good) translation. *arXiv preprint arXiv:1707.01626*.
- Adrien, G., Cécile, F., & Hakim, H. (2013). Predicting the temporal dynamics of information diffusion in social networks. *arXiv preprint arXiv:1302.5235*.
- Adrover, C., Bodnar, T., Huang, Z., Telenti, A., & Salathé, M. (2015). Identifying adverse effects of HIV drug treatment and associated sentiments using Twitter. *Journal of JMIR Public Health and Surveillance*, 1(2), e7.
- Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Languages in Social Media* (pp. 30-38). Association for Computational Linguistics, Portland, Oregon, US.
- Aghababaei, S., & Makrehchi, M. (2015). Temporal topic inference for trend prediction. In *2015 IEEE International Conference on Data Mining Workshop* (pp. 887-884). IEEE, Atlantic City, New Jersey, US.
- Aghababaei, S., & Makrehchi, M. (2017). Activity-based Twitter sampling for content-based and user-centric prediction models. *Journal of Human-Centric Computing and Information Sciences*, 7(3), 1-20.
- Al Hamoud, A., Alwehaibi, A., Roy, K., & Bikdash, M. (2018). Classifying political tweets using Naïve Bayes and Support Vector Machine. In *31st International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems* (pp. 736-744). Springer, Montreal, Quebec, Canada.
- Alp, Z. Z., & Öğüdücü, Ş. G. (2018). Influence factorization for identifying authorities in Twitter. *Journal of Knowledge-Based Systems*, 163, 944-954.
- Amoroso, V. N. A., Cham, N. A. C., Cruz, P. M. C., Monsale, C. R. E., San Jose, M. G. T., & Opiniano, J. M. (2018). #Trending: A reevaluation of traditional news values given Twitter through a mixed methods approach. *Jurnal Komunikasi: Malaysian Journal of Communication*, 34(2), 1-22.
- Anderson, A. S. M. (2018). Social media use in 2018. Retrieved on February 26th, 2019 from <http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/>.
- Anduiza, E., Cristancho, C., & Sabucedo, J. M. (2014). Mobilization through online social networks: The political protest of the Indignados in Spain. *Journal of Information, Communication & Society*, 17(6), 750-764.
- Anuta, D., Churchin, J., & Luo, J. (2017). Election bias: Comparing polls and Twitter in the 2016 US election. *arXiv preprint arXiv:1701.06232*.
- Araujo, M., Reis, J., Pereira, A., & Benevenuto, F. (2016). An evaluation of machine translation for multilingual sentence-level sentiment analysis. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing* (pp. 1140-1145). ACM, Pisa, Italy.
- Asher, M., Leston-Bandeira, C., & Spaiser, V. (2019). Do parliamentary debates of e-petitions enhance public engagement with Parliament? An analysis of Twitter conversations. *Journal of Policy & Internet*, 11(2), 149-171.
- Ashktorab, Z., Brown, C., Nandi, M., & Culotta, A. (2014). Tweedr: Mining Twitter to inform disaster response. In *Proceedings of the Conference on the International Association for Information Systems for Crisis Response and Management* (pp. 1-5). Pennsylvania, US.
- Auvinen, A.-M. (2012). Social media: The new power of political influence: Suomen Toivo Think Tank. Centre for European Studies.
- Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of*

- the 2010 Conference on Language Resource and Evaluation (pp. 2200-2204). European Languages Resources Association, Valletta, Malta.
- Bakliwal, A., Foster, J., van der Puil, J., O'Brien, R., Tounsi, L., & Hughes, M. (2013). Sentiment analysis of political tweets: Towards an accurate classifier. In *Proceedings of the Workshop on Language in Social Media* (pp. 49-58). Association for Computational Linguistics, Atlanta, Georgia, US.
- Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone is an influencer: Quantifying influence on Twitter. In *Proceedings of the International Conference on Web Search and Data Mining* (pp. 65-74). ACM, Hong Kong.
- Balahur, A. (2013). Sentiment analysis in social media texts. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis* (pp. 120-128). Association for Computational Linguistics, Atlanta, Georgia, US.
- Balahur, A., Mihalcea, R., & Montoyo, A. (2014). Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications. *Journal of Computer Speech and Language*, 28(2014), 1-6.
- Balahur, A., & Turchi, M. (2012). Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis* (pp. 52-60). Association for Computational Linguistics, Jeju, Republic of Korea.
- Baltas, A., Kanavos, A., & Tsakalidis, A. K. (2016). An Apache Spark implementation for sentiment analysis on Twitter data. In *International Workshop of Algorithmic Aspects of Cloud Computing* (pp. 15-25). Springer, Vienna, Austria.
- Bavaresco, A. (2014). Epistemology of social networks, public opinion and theory of agenda. *Journal of Canadian Social Science*, 10(6), 11-19.
- Bermingham, A., & Smeaton, A. (2011). On using Twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis: Where AI meets Psychology* (pp. 2-10). Workshop at the International Joint Conference for Natural Language Processing. Association for Computational Linguistics, Chiang Mai, Thailand.
- Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 US Presidential election online discussion. *Journal of Social Science Research Network*, 21(11), 1-14.
- Bhadane, C., Dalal, H., & Doshi, H. (2015). Sentiment analysis: Measuring opinions. *Procedia Computer Science*, 45(2015), 808-814.
- Bitecofer, R. (2018). Donald J. Trump: The making of a media event. *The unprecedented 2016 Presidential election* (pp. 39-58): Palgrave Macmillan, Cham.
- Blank, G. (2017). The digital divide among Twitter users and its implications for social research. *Journal of Social Science Computer Review*, 35(6), 679-697.
- Blight, M. G., Ruppel, E. K., & Schoenbauer, K. V. (2017). Sense of community on Twitter and Instagram: Exploring the roles of motives and parasocial relationships. *Journal of Cyberpsychology, Behavior, and Social Networking*, 20(5), 314-319.
- Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1-8.
- Boutet, A., & Yoneki, E. (2011). Member classification and party characteristics in Twitter during UK election. In *Proceedings of the 1st International Workshop on Dynamicity* (pp. 18-21). DYNAM, Toulouse, France.

- Bradshaw, S., & Howard, P. (2017). Troops, trolls and troublemakers: A global inventory of organized social media manipulation. *Oxford Internet Institute*, 12(2017), 1-37.
- Bradshaw, S., & Howard, P. (2018). Challenging truth and trust: A global inventory of organized social media manipulation: The Computational Propaganda Research Project. Oxford Internet Institute.
- Brena, G., Brambilla, M., Ceri, S., Di Giovanni, M., Pierri, F., & Ramponi, G. (2019). News sharing user behaviour on Twitter: A comprehensive data collection of news articles and social interactions. In *Proceedings of the International AAAI Conference on Web and Social Media* (pp. 592-597). AAAI Press, Atlanta, US.
- Brooks, M. (2015). *Human centered tools for analyzing online social data*. University of Washington.
- Brown, J., Broderick, A. J., & Lee, N. (2007). Word of mouth communication within online communities: Conceptualizing the online social network. *Journal of Interactive Marketing*, 21(3), 2-20.
- Brown, S., Hillegeist, S. A., & Lo, K. (2004). Conference calls and information asymmetry. *Journal of Accounting and Economics*, 37(3), 343-366.
- Bruns, A., & Stieglitz, S. (2013). Towards more systematic Twitter analysis: Metrics for tweeting activities. *International Journal of Social Research Methodology*, 16(2), 91-108.
- Budiharto, W., & Meiliana, M. (2018). Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis. *Journal of Big Data*, 5(1), 51-61.
- Burgess, J., & Bruns, A. (2012). (Not) The Twitter election: The dynamics of the #ausvotes conversation in relation to the Australian media ecology. *Journalism Practice*, 6(3), 384-402.
- Burnap, P., Gibson, R., Sloan, L., Southern, R., & Williams, M. (2016). 140 characters to victory? Using Twitter to predict the UK 2015 General election. *Journal of Electoral Studies*, 41(2016), 230-233.
- Burnap, P., & Williams, M. L. (2015). Cyber hate speech on Twitter: An application of machine classification and statistical modelling for policy and decision making. *Journal of Policy & Internet*, 7(2), 223-242.
- Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring user influence in Twitter: The million follower fallacy. In *Proceedings of the Fourth International AAAI Conference on Web and Social Media* (pp. 10-18). AAAI Press, Menlo Park, California, US.
- Chan, C.-H., & Fu, K.-W. (2017). The relationship between cyberbalkanization and opinion polarization: Time-series analysis on Facebook pages and opinion polls during the Hong Kong Occupy Movement and the associated debate on political reform. *Journal of Computer-Mediated Communication*, 22(5), 266-283.
- Chen, C., Zhang, J., Chen, X., Xiang, Y., & Zhou, W. (2015a). Six million spam tweets: A large ground truth for timely Twitter spam detection. In *Proceedings of the 2015 IEEE International Conference on Communications* (pp. 7065-7070). IEEE, London, UK.
- Chen, J., & Shen, X.-L. (2015). Consumers' decisions in social commerce context: An empirical investigation. *Journal of Decision Support Systems*, 79(2015), 55-64.
- Chen, X., Cho, Y., & Jang, S. Y. (2015b). Crime prediction using Twitter sentiment and weather. In *International Conference on Systems and Information*

- Engineering Design Symposium* (pp. 63-68). IEEE, Charlottesville, Virginia, US.
- Chepurna, I., Aghababaei, S., & Makrehchi, M. (2015). How to predict social trends by mining user sentiments. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (pp. 270-275). Springer, Cham, Washington, US.
- Chowdhury, S. G., Routh, S., & Chakrabarti, S. (2014). News analytics and sentiment analysis to predict stock price trends. *International Journal of Computer Science and Information Technologies*, 5(3), 3595-3604.
- Choy, M., Cheong, M. L., Laik, M. N., & Shung, K. P. (2011). A sentiment analysis of Singapore Presidential election 2011 using Twitter data with census correction. *arXiv preprint arXiv:1108.5520*.
- Cogburn, D. L., & Espinoza-Vasquez, F. K. (2011). From networked nominee to networked nation: Examining the impact of Web 2.0 and social media on political participation and civic engagement in the 2008 Obama campaign. *Journal of Political Marketing*, 10(1-2), 189-213.
- Coleman, S., & Gotze, J. (2001). *Bowling together: Online public engagement in policy deliberation*: Hansard Society London.
- Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. In *Proceedings of the Conference on Association for Information Science and Technology* (pp. 1-4). Association for Information Science and Technology, St. Louis, Missouri, US.
- Correa, J. C., & Camargo, J. E. (2017). Ideological consumerism in Colombian elections, 2015: Links between political ideology, Twitter activity, and electoral results. *Journal of Cyberpsychology, Behavior, and Social Networking*, 20(1), 37-43.
- Cottam, M. L., Mastors, E., Preston, T., & Dietz, B. (2015). *Introduction to political psychology*: Routledge.
- Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2015). Fame for sale: Efficient detection of fake Twitter followers. *Journal of Decision Support Systems*, 80(2015), 56-71.
- Culnan, M. J., McHugh, P. J., & Zubillaga, J. I. (2010). How large US companies can use Twitter and other social media to gain business value. *Journal of MIS Quarterly Executive*, 9(4), 243-259.
- Cwalina, W., Falkowski, A., & Newman, B. I. (2010). Towards the development of a cross-cultural model of voter behavior: Comparative analysis of Poland and the US. *European Journal of Marketing*, 44(3-4), 351-368.
- Das, A., & Bandyopadhyay, S. (2010). SentiWordNet for Indian languages. In *Proceedings of the Eighth Workshop on Asian Language Resources* (pp. 56-63). Asian Federation for Natural Language Processing, Beijing, China.
- Davies, C. (2015). GE 2015: Can electoral decision making be rational? Retrieved on April 5th, 2016 from <http://policybristol.blogs.bris.ac.uk/2015/05/07/ge-2015-can-electoral-decision-making-be-rational/>.
- De Marneffe, M.-C., MacCartney, B., & Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the International Conference on Language Resources and Evaluation* (pp. 449-454). European Language Resources Association, Genova, Italy.
- Di Bari, M., Sharoff, S., & Thomas, M. (2015). A manually-annotated Italian corpus for fine-grained sentiment analysis. In *Proceedings of the Second Italian Conference on Computational Linguistics* (pp. 105-109). Accademia University Press, Torino, Italy.

- DiGrazia, J., McKelvey, K., Bollen, J., & Rojas, F. (2013). More tweets, more votes: Social media as a quantitative indicator of political behavior. *Journal of PLOS ONE*, 8(11), e79449.
- Dwi Prasetyo, N., & Hauff, C. (2015). Twitter-based election prediction in the developing world. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media* (pp. 149-158). ACM, New York, US.
- Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., . . . Sap, M. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Journal of Psychological Science*, 26(2), 159-169.
- Ensley, M. J. (2007). Candidate divergence, ideology, and vote choice in US Senate elections. *Journal of American Politics Research*, 35(1), 103-122.
- Eriksson, M. (2016). Managing collective trauma on social media: The role of Twitter after the 2011 Norway attacks. *Journal of Media, Culture & Society*, 38(3), 365-380.
- Fan, W., & Gordon, M. D. (2014). The power of social media analytics. *Journal of Communications of the ACM*, 57(6), 74-81.
- Ferrara, E. (2017). Disinformation and social bot operations in the run up to the 2017 French Presidential election. *Journal of First Monday*, 22(8), 1-33.
- Fountaine, S. (2017). What's not to like? A qualitative study of young women politicians' self-framing on Twitter. *Journal of Public Relations Research*, 29(5), 219-237.
- George, G., Haas, M. R., & Pentland, A. (2014). *Big data and management*: Academy of Management Briarcliff Manor, NY.
- Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., . . . Smith, N. A. (2010). Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (pp. 42-47). Association for Computational Linguistics, Portland, Oregon, US.
- Go, A., Huang, L., & Bhayani, R. (2009). Twitter sentiment analysis. *Journal of Entropy*, 17, 252-258.
- Goodrich, K., & De Mooij, M. (2014). How "social" are social media? A cross-cultural comparison of online and offline purchase decision influences. *Journal of Marketing Communications*, 20(1-2), 103-116.
- Gorodnichenko, Y., Pham, T., & Talavera, O. (2018). Social media, sentiment and public opinions: Evidence from #Brexit and #USElection: National Bureau of Economic Research.
- Grover, P., Kar, A. K., Dwivedi, Y. K., & Janssen, M. (2019). Polarization and acculturation in US Election 2016 outcomes. Can Twitter analytics predict changes in voting preferences? *Journal of Technological Forecasting and Social Change*, 145, 438-460.
- Gulati, G. J., & Williams, C. B. (2015). Congressional campaigns' motivations for social media adoption. In V. A. Farrar-Myers, & J. S. Vaughn (Eds.), *Controlling the message: New media in American political campaigns* (pp. 32-52). New York: NYU Press.
- Gupta, A., Lamba, H., & Kumaraguru, P. (2013). \$1.00 per RT #BostonMarathon #PrayForBoston: Analyzing fake content on Twitter. In *Proceedings of the 2013 eCrime Researchers Summit* (pp. 1-12). IEEE, Delhi, India.
- Haddi, E., Liu, X., & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Journal of Procedia Computer Science*, 17(2013), 26-32.

- Halberstam, Y., & Knight, B. (2016). Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. *Journal of Public Economics*, 143, 73-88.
- Henderson, J., Wilson, A. M., Webb, T., McCullum, D., Meyer, S. B., Coveney, J., & Ward, P. R. (2017). The role of social media in communication about food risks: Views of journalists, food regulators and the food industry. *British Food Journal*, 119(3), 453-467.
- Hoang, T. B. N., & Mothe, J. (2017). Predicting information diffusion on Twitter. Analysis of predictive features. *Journal of Computational Science*, 28(2017), 257-264.
- Hoffman, D. L., & Fodor, M. (2010). Can you measure the ROI of your social media marketing? *MIT Sloan Management Review*, 52(1), 41-49.
- Hong, L., Doumith, A. S., & Davison, B. D. (2013). Co-factorization machines: Modelling user interests and predicting individual decisions in Twitter. In *Proceedings of the ACM International Conference on Web Search and Data Mining* (pp. 557-566). ACM, Rome, Italy.
- Howard, P. N., & Kollanyi, B. (2016). Bots, #StrongerIn, and #Brexit: Computational propaganda during the UK-EU referendum. *Oxford Internet Institute*. Working Paper 2016.1: The Computational Propaganda Project.
- Hsu, C.-l., & Park, H. W. (2012). Mapping online social networks of Korean politicians. *Journal of Government Information Quarterly*, 29(2), 169-181.
- Hu, H., Wen, Y., Chua, T.-S., & Li, X. (2014). Toward scalable systems for Big data analytics: A technology tutorial. *Journal of IEEE Access*, 2, 652-687.
- Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM International Conference on Knowledge Discovery and Data Mining* (pp. 168-177). ACM, Seattle, Washington, US.
- Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International Conference on Weblogs and Social Media* (pp. 216-226). AAAI Press, Ann Arbor, Michigan, US.
- Ibrahim, N. F., Wang, X., & Bourne, H. (2017). Exploring the effect of user engagement in online brand communities: Evidence from Twitter. *Journal of Computers in Human Behavior*, 72, 321-338.
- Ince, J., Rojas, F., & Davis, C. A. (2017). The social media response to Black Lives Matter: how Twitter users interact with Black Lives Matter through hashtag use. *Journal of Ethnic and racial studies*, 40(11), 1814-1830.
- Innes, M., Roberts, C., Preece, A., & Rogers, D. (2018). Ten “Rs” of social reaction: Using social media to analyse the “post-event” impacts of the murder of Lee Rigby. *Journal of Terrorism and Political Violence*, 30(3), 454-474.
- Jain, A., & Jain, V. (2019). Sentiment classification of Twitter data belonging to renewable energy using machine learning. *Journal of Information and Optimization Sciences*, 40(2), 521-533.
- Jain, V. K., & Kumar, S. (2017). Towards prediction of election outcomes using social media. *International Journal of Intelligent Systems and Applications*, 9(12), 20-28.
- Jin, S.-A. A., & Phua, J. (2014). Following celebrities’ tweets about brands: The impact of Twitter-based electronic word-of-mouth on consumers’ source credibility perception, buying intention, and social identification with celebrities. *Journal of Advertising*, 43(2), 181-195.
- Joachims, T. (1999). Making large-scale SVM learning practical. *Advances in kernel methods*: MIT Press, Cambridge.

- Jose, R., & Chooralil, V. S. (2015). Prediction of election result by enhanced sentiment analysis on Twitter data using word sense disambiguation. In *2015 International Conference on Control Communication & Computing* (pp. 638-641). IEEE, Trivandrum, India.
- Joshi, A., Balamurali, A., & Bhattacharyya, P. (2010). A fall-back strategy for sentiment analysis in Hindi: A case study. In *Proceedings of the 8th International Conference on Natural Language Processing* (pp. 1-6. Macmillan Publishers, Kharagpur, India.
- Jost, J. T., Barberá, P., Bonneau, R., Langer, M., Metzger, M., Nagler, J., . . . Tucker, J. A. (2018). How social media facilitates political protest: Information, motivation, and social networks. *Journal of Political Psychology*, 39, 85-118.
- Juris, J. S. (2012). Reflections on #Occupy Everywhere: Social media, public space, and emerging logics of aggregation. *Journal of American Ethnologist*, 39(2), 259-279.
- Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big data: Issues and challenges moving forward. In *6th Hawaii International Conference on System Sciences* (pp. 995-1004). IEEE, Hawaii, US.
- Kalimeri, K., Beiró, M. G., Bonanomi, A., Rosina, A., & Cattuto, C. (2019). Evaluation of biases in self-reported demographic and psychometric information: Traditional versus Facebook-based surveys. *arXiv preprint arXiv:1901.07876*.
- Karimi, F., Génois, M., Wagner, C., Singer, P., & Strohmaier, M. (2018). Homophily influences ranking of minorities in social networks. *Journal of Scientific Reports*, 8, 1-12.
- Karlsen, R. (2013). Obama's online success and European party organizations: Adoption and adaptation of US online practices in the Norwegian labour party. *Journal of Information Technology & Politics*, 10(2), 158-170.
- Karusala, N., Kumar, N., & Arriaga, R. (2019). #Autism: Twitter as a lens to explore differences in autism awareness in India and the United States. In *Proceedings of the International Conference on Information and Communication Technologies, and Development* (pp. 1-5). ACM, Gujarat, India.
- Kazimieras Zavadskas, E., Antucheviciene, J., & Chatterjee, P. (2019). Multiple-criteria decision making (MCDM) techniques for business processes information management: Multidisciplinary Digital Publishing Institute.
- Khan, F. H., Bashir, S., & Qamar, U. (2014). TOM: Twitter opinion mining framework using hybrid classification scheme. *Journal of Decision Support Systems*, 57, 245-257.
- Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the Annual Meeting on Association for Computational Linguistics* (pp. 423-430). Association for Computational Linguistics, Sapporo, Japan.
- Klein, G. A. (2017). *Sources of power: How people make decisions*: MIT press.
- Knobloch-Westerwick, S., Johnson, B. K., & Westerwick, A. (2014). Confirmation bias in online searches: Impacts of selective exposure before an election on political attitude strength and shifts. *Journal of Computer-Mediated Communication*, 20(2), 171-187.
- Komorowski, M., Do Huu, T., & Deligiannis, N. (2018). Twitter data analysis for studying communities of practice in the media industry. *Journal of Telematics and Informatics*, 35(1), 195-212.
- Kruikemeier, S. (2014). How political candidates use Twitter and the impact on votes. *Journal of Computers in Human Behavior*, 34, 131-139.

- Kurka, D. B., Godoy, A., & Von Zuben, F. J. (2017). Using retweet information as a feature to classify messages contents. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 1485-1491). International World Wide Web Conferences Steering Committee, Geneva, Switzerland.
- Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media?. In *Proceedings of the 19th International Conference on World Wide Web* (pp. 591-600). ACM, Raleigh, North Caroline, US.
- Lachat, R. (2008). The impact of party polarization on ideological voting. *Journal of Electoral Studies*, 27(4), 687-698.
- Larsson, A. O., & Kalsnes, B. (2014). "Of course we are on Facebook": Use and non-use of social media among Swedish and Norwegian politicians. *European Journal of Communication*, 29(6), 653-667.
- Lassen, N. B., la Cour, L., & Vatrapu, R. (2017). Predictive analytics with social media data. In L. Sloan & A. Quan-Haase (Eds.), *The SAGE Handbook of Social Media Research Methods* (pp. 328-341): SAGE.
- Lau, R. R. (2003). Models of decision making. *Oxford Handbook of Political Psychology* (pp. 19-59): Sears, Huddy and Jervis.
- Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). *Mining of massive datasets* (Second ed.): Cambridge University Press.
- Lin, Y.-R., Bagrow, J. P., & Lazer, D. (2011). More voices than ever? Quantifying media bias in networks. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (pp. 193-200). AAAI Press, Barcelona, Spain.
- Littman, J., Wrubel, L., & Kerchner, D. (2016). 2016 United States presidential election tweet IDs: Harvard Dataverse.
- Lloyd, A., & Cheshire, J. (2016). Mining consumer insights from geo-located social media datasets. In *Proceedings of the GIS Research UK Conference* (pp. 1-7). GIS Research UK, London, UK.
- Lohar, P., Afli, H., & Way, A. (2017). Maintaining sentiment polarity in translation of user-generated content. *The Prague Bulletin of Mathematical Linguistics*, 108(1), 73-84.
- Loria, S., Keen, P., Honnibal, M., Yankovsky, R., Karesh, D., & Dempsey, E. (2014). Textblob: Simplified text processing. *TextBlob*. Retrieved on February 12th, 2019 from <https://textblob.readthedocs.io/en/dev/index.html>.
- Makazhanov, A., Rafiei, D., & Waqar, M. (2014). Predicting political preference of Twitter users. In *Proceedings of the 2013 International Conference on Advances in Social Networks Analysis and Mining* (pp. 298-305). IEEE, Niagara Falls, Ontario, Canada.
- Malik, M. M., Lamba, H., Nakos, C., & Pfeffer, J. (2015). Population bias in geotagged tweets. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media* (pp. 18-28). AAAI Press, Oxford, UK.
- Malik, P. (2013). Governing Big data: Principles and practices. *IBM Journal of Research and Development*, 57(3-4), 1:1-1:13.
- Mandviwalla, M., & Watson, R. (2014). Generating capital from social media. *Journal of MIS Quarterly Executive*, 13(2), 97-113.
- Martin-Sanchez, F., & Verspoor, K. (2014). *Big data in medicine is driving big changes* (Vol. 9): Yearbook of Medical Informatics.
- McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., & Barton, D. (2012). Big data: The management revolution. *Harvard Business Review*, 90(10), 60-68.

- McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Journal of Annual Review of Sociology*, 27(1), 415-444.
- Mellon, J., & Prosser, C. (2017). Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Journal of Research & Politics*, 4(3), 1-9.
- Meyer, D. (2018). Twitter under formal investigation for how it tracks users in the GDPR era. Retrieved on February 20th, 2019 from <http://fortune.com/2018/10/12/twitter-gdpr-investigation-tco-tracking/>.
- Mihalcea, A.-D., & Savulescu, R.-M. (2013). Social networking sites: Guidelines for creating new business opportunities through Facebook, Twitter, and LinkedIn. *Management Dynamics in the Knowledge Economy*, 1(1), 39-53.
- Moghaddam, S. (2015). Beyond sentiment analysis: Mining defects and improvements from customer feedback. In *European Conference on Advances in Information Retrieval* (pp. 400-410). Springer, Vienna, Austria.
- Morozov, E. (2011). The net delusion: The dark side of Internet freedom. *New York: Public Affairs*.
- Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media* (pp. 400-409). AAAI Press, Cambridge, Massachusetts, US.
- Munir, S. (2018). Social Media and shaping voting behavior of youth: The Scottish Referendum 2014 case. *The Journal of Social Media in Society*, 7(1), 253-279.
- Nesi, P., Pantaleo, G., Paoli, I., & Zaza, I. (2018). Assessing the retweet proneness of tweets: Predictive models for retweeting. *Journal of Multimedia Tools and Applications*, 77(20), 26371–26396.
- Novak, M. (2018). Facebook and Google accused of violating GDPR on first day of the new European privacy law. Retrieved on February 20th, 2019 from <https://gizmodo.com/facebook-and-google-accused-of-violating-gdpr-on-first-1826321323>.
- Nusko, B., Tahmasebi, N., & Mogren, O. (2016). Building a sentiment lexicon for Swedish. In *Proceedings of the Workshop of Digitization to Knowledge: Resources and Methods for Semantic Processing of Digital Works/Texts* (pp. 32-37). Linköping University Electronic Press, Krakow, Poland.
- O'Connor, B., Krieger, M., & Ahn, D. (2010). TweetMotif: Exploratory search and topic summarization for Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media* (pp. 384-385). AAAI Press, Washington, US.
- Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *International Journal of Advanced Research in Computer and Communication Engineering*, 10, 1320-1326.
- Panagiotopoulos, P., Bigdeli, A. Z., & Sams, S. (2014). Citizen - Government collaboration on social media: The case of Twitter in the 2011 riots in England. *Journal of Government Information Quarterly*, 31(3), 349-357.
- Pandarachalil, R., Sendhilkumar, S., & Mahalakshmi, G. (2015). Twitter sentiment analysis for large-scale data: An unsupervised approach. *Journal of Cognitive Computation*, 7(2), 254-262.
- Pandey, P., Kumar, M., & Srivastava, P. (2016). Classification techniques for Big data: A survey. In *3rd International Conference on Computing for Sustainable Global Development* (pp. 3625-3629). IEEE, New Delhi, India.

- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Journal of Foundations and Trends in Information Retrieval*, 2(1–2), 1-135.
- Pearce, K. E., & Kendzior, S. (2012). Networked authoritarianism and social media in Azerbaijan. *Journal of Communication*, 62(2), 283-298.
- Penlington, R. (2017). Predicting human behaviour with Big data. Retrieved on March 5th, 2019 from <https://www.linkedin.com/pulse/predicting-human-behavior-big-data-russ-penlington/>.
- Pennacchiotti, M., & Popescu, A.-M. (2011a). Democrats, Republicans and Starbucks aficionados: User classification in Twitter. In *Proceedings of the 17th ACM International Conference on Knowledge Discovery and Data Mining* (pp. 430-438). ACM, San Diego, California, US.
- Pennacchiotti, M., & Popescu, A.-M. (2011b). A machine learning approach to Twitter user classification. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (pp. 281-288). AAAI Press, Barcelona, Spain.
- Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic Inquiry and Word Count: LIWC2007.
- Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. *The University of Texas at Austin*.
- Persily, N. (2017). The 2016 US election: Can democracy survive the Internet? *Journal of Democracy*, 28(2), 63-76.
- Pons, V. (2016). Has social science taken over electoral campaigns and should we regret it?. *Journal of French Politics, Culture & Society*, 34(1), 34-47.
- Poortinga, W., & Pidgeon, N. F. (2004). Trust, the asymmetry principle, and the role of prior beliefs. *Journal of Risk Analysis*, 24(6), 1475-1486.
- Preoțiuc-Pietro, D., Liu, Y., Hopkins, D., & Ungar, L. (2017). Beyond binary labels: Political ideology prediction of Twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (pp. 729-740). Association for Computational Linguistics, Vancouver, Canada.
- Procter, R., Vis, F., & Voss, A. (2013). Reading the riots on Twitter: Methodological innovation for the analysis of Big data. *International Journal of Social Research Methodology*, 16(3), 197-214.
- Punit, I. S. (2018). 335 Indians installed a Cambridge Analytica app, exposing the Facebook data of 560,000. *Quartz India*. Retrieved on September 16th, 2018 from <https://qz.com/india/1245515/facebook-admits-cambridge-analytica-may-have-accessed-the-data-of-over-560000-users-in-india/>.
- Python. (2019): Python Software Foundation. Python Language Reference, version 3.5.7. Available at <http://www.python.org/>.
- R Core Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (Version 1.0.136). Retrieved from <http://www.r-project.org/>.
- Ramteke, J., Shah, S., Godhia, D., & Shaikh, A. (2016). Election result prediction using Twitter sentiment analysis. In *Proceedings of the International Conference on Inventive Computation Technologies* (pp. 1-5). IEEE, Tamilnadu, India.
- Rathan, M., Hulipalled, V. R., Venugopal, K., & Patnaik, L. (2018). Consumer insight mining: Aspect based Twitter opinion mining of mobile phone reviews. *Journal of Applied Soft Computing*, 68, 765-773.
- Redlawsk, D. P. (2004). What voters do: Information search during election campaigns. *Journal of Political Psychology*, 25(4), 595-610.

- Reed, M. (2015). Social network influence on consistent choice. *Journal of Choice Modelling*, 17, 28-38.
- Remus, R., Quasthoff, U., & Heyer, G. (2010). SentiWS: A publicly available German-language resource for sentiment analysis. In *Proceedings of the International Conference on Language Resources and Evaluation* (pp. 1168-1171). European Languages Resources Association, Valletta, Malta.
- Reuter, O. J., & Szakonyi, D. (2015). Online social media and political awareness in authoritarian regimes. *British Journal of Political Science*, 45(1), 29-51.
- Revelli, F. (2002). Local taxes, national politics and spatial interactions in English district election results. *European Journal of Political Economy*, 18(2), 281-299.
- Ribeiro, S. S., Davis Jr, C. A., Oliveira, D. R. R., Meira Jr, W., Gonçalves, T. S., & Pappa, G. L. (2012). Traffic observatory: A system to detect and locate traffic events and conditions using Twitter. In *Proceedings of the ACM International Workshop on Location-Based Social Networks* (pp. 5-11). ACM, Redondo Beach, California, US.
- Robertson, J., Riley, M., & Willis, A. (2016). How to hack an election. *Bloomberg*. Retrieved on June 7th, 2016 from <http://www.bloomberg.com/features/2016-how-to-hack-an-election/>.
- Rolls, E. T., Warwick, C., Bilderbeck, A., Bowtell, R., Browning, A., Critchley, H., . . . Hornak, J. (2019). Emotion and reasoning in human decision making. *Journal of Economics*, 2, 1-20.
- Rost, M., Barkhuus, L., Cramer, H., & Brown, B. (2013). Representation and communication: Challenges in interpreting large social media datasets. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (pp. 357-362). ACM, San Antonio, Texas, US.
- Ruba, K. V., & Venkatesan, D. (2015). Building a custom sentiment analysis tool based on an ontology for Twitter posts. *Indian Journal of Science and Technology*, 8(13), 1-9.
- Rubin, V., Conroy, N., Chen, Y., & Cornwell, S. (2016). Fake news or truth? Using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection* (pp. 7-17). Association for Computational Linguistics, San Diego, California, US.
- Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. In *2013 International Conference on Collaboration Technologies and Systems* (pp. 42-47). IEEE, San Diego, California, US.
- Sang, E. T. K., & Bos, J. (2012). Predicting the 2011 Dutch Senate election results with Twitter. In *Proceedings of the Workshop on Semantic Analysis in Social Media* (pp. 53-60). Association for Computational Linguistics, Stroudsburg, Pennsylvania, US.
- Saravanakumar, M., & SuganthaLakshmi, T. (2012). Social media marketing. *Life Science Journal*, 9(4), 4444-4451.
- Sarewitz, D., Pielke, R. A., & Byerly, R. (2000). Prediction in science and policy. *Journal of Technology in Society*, 21(2), 11-22.
- Schipper, B. C., & Woo, H. (2017). Political awareness, microtargeting of voters, and negative electoral campaigning. *Journal of SSRN*, May 2, 2017.
- Schofield, P., & Reeves, P. (2015). Does the factor theory of satisfaction explain political voting behaviour? *European Journal of Marketing*, 49(5-6), 968-992.
- Sharma, B., & Parma, S. (2016). Impact of social media on voter's behaviour. A descriptive study of Gwalior, Madhya Pradesh. *International Journal of Research in Computer Science and Management*, 4(1), 5-8.

- Sharma, P., & Moh, T.-S. (2016). Prediction of Indian election using sentiment analysis on Hindi Twitter. In *International Conference on Big Data* (pp. 1966-1971). IEEE, Washington, US.
- Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *Journal of MIS Quarterly*, 35(3), 553-572.
- Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *Journal of ACM SIGKDD Explorations*, 19(1), 22-36.
- Simon, T., Goldberg, A., Aharonson-Daniel, L., Leykin, D., & Adini, B. (2014). Twitter in the cross fire: The use of social media in the Westgate Mall terror attack in Kenya. *Journal of PLOS ONE*, 9(8), e104136.
- Sinnenberg, L., Buttenheim, A. M., Padrez, K., Mancheno, C., Ungar, L., & Merchant, R. M. (2017). Twitter as a tool for health research: A systematic review. *American Journal of Public Health*, 107(1), e1-e8.
- Skoric, M., Poor, N., Achananuparp, P., Lim, E.-P., & Jiang, J. (2012). Tweets and votes: A study of the 2011 Singapore General election. In *45th Hawaii International Conference on System Sciences* (pp. 2583-2591). IEEE, Maui, Hawaii, US.
- Smailović, J., Grčar, M., Lavrač, N., & Žnidaršič, M. (2013). Predictive sentiment analysis of tweets: A stock market application. *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data* (pp. 77-88): Springer.
- Song, M., & Kim, M. C. (2013). RT²M: Real-time Twitter trend mining system. In *2013 International Conference on Social Intelligence and Technology* (pp. 64-71). IEEE, Pennsylvania, US.
- Spohr, D. (2017). Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Journal of Business Information Review*, 34(3), 150-160.
- Sprenger, T. O., Sandner, P. G., Tumasjan, A., & Welpe, I. M. (2014). News or noise? Using Twitter to identify and understand company-specific news flow. *Journal of Business Finance & Accounting*, 41(7-8), 791-830.
- Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010). Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In *2010 IEEE International Conference on Social Computing* (pp. 177-184). IEEE, Minneapolis, Minnesota, US.
- Sun, S., & Rui, H. (2017). Link formation on Twitter: The role of achieved status and value homophily. In *Proceedings of the 50th Hawaii International Conference on System Sciences* (pp. 5609-5618). Hawaii, US.
- Sunstein, C. R. (2009). *Republic.com 2.0*: Princeton University Press.
- Suthaharan, S. (2014). Big data classification: Problems and challenges in network intrusion prediction with machine learning. *ACM SIGMETRICS Performance Evaluation Review*, 41(4), 70-73.
- Tamine, L., Soulier, L., Ben Jabeur, L., Amblard, F., Hanachi, C., Hubert, G., & Roth, C. (2016). Social media-based collaborative information access: Analysis of online crisis-related Twitter conversations. In *Proceedings of the 27th ACM Conference on Hypertext and Social Media* (pp. 159-168). ACM, Halifax, Nova Scotia, Canada.
- Thackeray, R., Neiger, B. L., Hanson, C. L., & McKenzie, J. F. (2008). Enhancing promotional strategies within social marketing programs: Use of Web 2.0 social media. *Journal of Health Promotion Practice*, 9(4), 338-343.

- Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12), 2544-2558.
- Tuarob, S., & Tucker, C. S. (2013). Fad or here to stay: Predicting product market adoption and longevity using large scale, social media data. In *Proceedings of the ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* (pp. 1-13). Volume 2B: 33rd Computers and Information in Engineering Conference. Portland, Oregon, US.
- Tufekci, Z., & Wilson, C. (2012). Social media and the decision to participate in political protest: Observations from Tahrir Square. *Journal of Communication*, 62(2), 363-379.
- Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. *Journal of The International Linguistic Association*, 10(1), 178-185.
- Tunggawan, E., & Soelistio, Y. E. (2016). And the winner is...: Bayesian Twitter-based prediction on 2016 US Presidential election. In *2016 International Conference on Computer, Control, Informatics, and its Applications* (pp. 33-37). IEEE, Tangerang, Indonesia.
- Valenzuela, S. (2013). Unpacking the use of social media for protest behavior: The roles of information, opinion expression, and activism. *Journal of American Behavioral Scientist*, 57(7), 920-942.
- Valenzuela, S., Arriagada, A., & Scherman, A. (2012). The social media basis of youth protest behavior: The case of Chile. *Journal of Communication*, 62(2), 299-314.
- Vatrapu, R., Hussain, A., Lassen, N. B., Mukkamala, R. R., Flesch, B., & Madsen, R. (2015). Social set analysis: Four demonstrative case studies. In *Proceedings of the 2015 International Conference on Social Media & Society* (pp. 1-9). ACM, Toronto, Canada.
- Verma, R., & Sardesai, S. (2014). Does media exposure affect voting behaviour and political preferences in India. *Journal of Economic and Political Weekly*, 49(39), 82-88.
- Vis, F. (2013). Twitter as a reporting tool for breaking news: Journalists tweeting the 2011 UK riots. *Journal of Digital Journalism*, 1(1), 27-47.
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Journal of Science*, 359(6380), 1146-1151.
- Vural, A. G., Cambazoglu, B. B., Senkul, P., & Tokgoz, Z. O. (2013). A framework for sentiment analysis in Turkish: Application to polarity detection of movie reviews in Turkish. *Computer and Information Sciences III* (pp. 437-445): Springer.
- Wang, A. H. (2010). Don't follow me: Spam detection in Twitter. In *2010 International Conference on Security and Cryptography* (pp. 1-10). IEEE, Athens, Greece.
- Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.
- Wang, X., Gerber, M. S., & Brown, D. E. (2012). Automatic crime prediction using events extracted from Twitter posts. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (pp. 231-238). Springer, College Park, Maryland, US.
- Webb, H., Jirotko, M., Stahl, B. C., Housley, W., Edwards, A., Williams, M., . . . Burnap, P. (2017). The ethical challenges of publishing Twitter data for

- research dissemination. In *Proceedings of the ACM International Conference on Web Science Conference* (pp. 339-348). ACM, Troy, New York, US.
- Weiler, R. (2013). *We the people: The role of social media in the participatory community of the Tea Party movement*. (MSc in Politics and Communication), University of London, London.
- Wicaksono, A. F., Vania, C., Distiawan, B., & Adriani, M. (2014). Automatically building a corpus for sentiment analysis on Indonesian tweets. In *Proceedings of the 28th Pacific Asia Conference on Language, Information, and Computing* (pp. 185-194). Department of Linguistics, Chulalongkorn University, Phuket, Thailand.
- Williams, M. L., Burnap, P., & Sloan, L. (2017). Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Journal of Sociology*, 51(6), 1149-1168.
- Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 347-354). Association for Computational Linguistics, Vancouver, British Columbia, Canada.
- Wirawanda, Y., & Wibowo, T. O. (2018). Twitter: Expressing hate speech behind tweeting. *Profetik: Jurnal Komunikasi*, 11(1), 5-11.
- Wu, X., Zhu, X., Wu, G.-Q., & Ding, W. (2014). Data mining with Big data. *Journal of IEEE Transactions on Knowledge and Data Engineering*, 26(1), 97-107.
- Yang, J., & Leskovec, J. (2011). Patterns of temporal variation in online media. In *Proceedings of the ACM International Conference on Web Search and Data Mining* (pp. 177-186). ACM, New York, US.
- Younus, A., Qureshi, M. A., Saeed, M., Touheed, N., O'Riordan, C., & Pasi, G. (2014). Election trolling: Analyzing sentiment in tweets during Pakistan elections 2013. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 411-412). ACM, New York, US.
- Zafar, M. B., Bhattacharya, P., Ganguly, N., Gummadi, K. P., & Ghosh, S. (2015). Sampling content from online social networks: Comparing random vs. expert sampling of the Twitter stream. *Journal of ACM Transactions on the Web*, 9(3), 12-33.
- Zhang, J., Tang, J., Li, J., Liu, Y., & Xing, C. (2015). Who influenced you? Predicting retweet via social influence locality. *Journal of ACM Transactions on Knowledge Discovery from Data*, 9(3), 1-27.
- Zhao, J. (2015). Pre-processing boosting Twitter sentiment analysis? In *2015 IEEE International Conference on Smart City* (pp. 748-753). IEEE, Chengdu, China.
- Zhao, J., & Gui, X. (2017). Comparison research on text pre-processing methods on Twitter sentiment analysis. *Journal of IEEE Access*, 5, 2870-2879.
- Zhao, J., Gui, X., & Zhang, X. (2018). Deep convolution neural networks for Twitter sentiment analysis. *Journal of IEEE Access*, 6, 23253-23260.
- Zimmer, M., & Proferes, N. J. (2014). A topology of Twitter research: Disciplines, methods, and ethics. *Aslib Journal of Information Management*, 66(3), 250-261.
- Zimper, A., & Ludwig, A. (2009). On attitude polarization under Bayesian learning with non-additive beliefs. *Journal of Risk and Uncertainty*, 39(2), 181-212.

Chapter 5

Making progress on Twitter data analysis: A study based on sentiment analysis and identification of influential users

Abstract: This study introduces two new approaches to data pre-processing and influential users' identification when conducting Twitter-based sentiment modelling, which addresses two overlooked aspects. First, the need to integrate features used in everyday tweeting that are often discarded in sentiment modelling, such as hashtags, emoticons, and URLs. This paper offers a coding approach to these features for readability to conduct sentiment analysis. And second, the identification of users that are momentarily influential in the sentiment distribution of tweets. The study used 1.3 million tweets generated during the 2017 Ecuadorian Presidential election that were relevant to the two most voted candidates. The results show that: first, the pre-processing approach enhances sentiment classification when compared to sentiment analysis using raw tweets; second, the candidates' accounts were consistently the most influential through the campaign, while other influential users changed every week; and third, the count of unique Twitter users producing positive sentiment towards a candidate was found to provide more accurate results of vote share than polling firms, with a sentiment detection error of 0.25% in the second round. Finally, this study offers academics and campaigners a new way of measuring vote share and ranking influencers affecting a candidate.

Keywords: Twitter, Sentiment analysis, Influential users, Data pre-processing, Politics.

Declaration of interest: None.

5.1. Introduction

This study introduces two new approaches to Twitter data analysis, which are respectively relevant to data pre-processing for sentiment classification and to influential user identification. Evidence suggests that Twitter analysis offers some advantages over traditional polling for studying a wide range of social dynamics (Bovet, Morone, & Makse, 2016; Martínez-Cámara, Martín-Valdivia, Urena-López, & Montejo-Ráez, 2014; Vidal-Alaball, Fernandez-Luque, Marin-Gomez, & Ahmed, 2019). In addition, unlike conducting traditional polls, which can be time-consuming

and often demand higher financial and labour resources, Twitter content can provide a picture of how values and meanings attached to users, goods, and services are shared and clustered among users in real-time and at low cost. Consequently, monitoring Twitter is fast becoming a regular practice and being positioned as an effective approach to make real-time social measurements of a different nature (Sharma, Rokne, & Alhajj, 2018; Weng, 2019). Moreover, private firms are also using Twitter as a core tool for brand-building strategies (Cawsey & Rowley, 2016; McShane, Pancer, & Poole, 2019). Often, the purpose is to identify drivers that propel user decisions because it can inform the designing of marketing campaigns.

In politics, Twitter became and continues to be a key platform for users to share ideas, concerns, and views on politicians and parties. Concerning politicians, they have adopted Twitter globally as a channel for self-promotion (Enli & Skogerbø, 2013; Golbeck, Grimes, & Rogers, 2010), for discussing personal and political issues (Larsson & Moe, 2012; Lee & Xu, 2018; Small, 2011), for positioning political discourses (Lee & Xu, 2018), and ultimately as an instrument for partisan and voter mobilisation (Linh, Stieglitz, Wladarsch, & Neuberger, 2013).

The use of Twitter gains relevance when looking for an alternative way of communication. This is especially valued when politicians believe traditional media are not providing them sufficient and fair coverage, so they become empowered to act themselves as media channels (Enli, 2017). Thus, politicians are aware of the value of Twitter to address their aims of influence, and thereafter increase their popularity and recognition. On the other hand, users often use Twitter as a platform to support or oppose politicians, while adopting a more active role (Enli & Skogerbø, 2013) that allows interaction with other users who support or challenge their viewpoints. It is this continuous generation of Twitter data that constitutes a powerful resource to develop models that challenge the convenience of conventional polls.

One of the most widespread analytical tools used when attempting to predict electoral outcomes with Twitter content is sentiment analysis. This consists of identifying and tagging positive and negative opinions or feelings from a text about a given entity (Nakov et al., 2016). On Twitter, an entity can refer to users, goods, services, or places. Sentiment analysis can include neutral feelings when a text contains both positive and negative sentiments, or none when no sentiment can be detected. A range of academics have become interested in determining the sentiment of users, with data classification algorithms such as decision trees, Naïve Bayes,

support vector machines, and others, to estimate their level of support to a given candidate for office, and to capture their positions about issues of political debate (Arcuri et al., 2008; Friese et al., 2012; Roccato & Zogmaister, 2010). However, challenges of accuracy and performance remain when applying algorithms to experimental data. About this, poor sentiment detection may sometimes be linked to the noisiness of Twitter data, which is linked to the presence of meaningless data that do not add value to the sentiment analysis. It is in these circumstances that data pre-processing becomes relevant, as addressing noisiness would improve overall outcomes.

In terms of identification of influential users, this is relevant because it helps in detecting who are setting the agenda of the political conversation. According to the Oxford dictionary, influence refers to “the capacity to have an effect on the character, development, or behaviour of someone or something, or the effect itself”¹. On Twitter, influence is usually tied to those users with the ability to engage others in public conversations, generate and distribute tweets widely retweeted, and integrate with a network of experts and followers within a specific market (Chandawarkar, Gould, & Grant Stevens, 2018; Shattell & Darmoc, 2017). The identification of influential users on Twitter has been historically carried out by analysing the distribution of networks using different tools. For example, social network analysis (SNA) is today a conventional approach to measure a user’s activity connections and estimate the ability to spread content (Cossu, Dugué, & Labatut, 2015; Kim et al., 2018; Maharani, Adiwijaya, & Gozali, 2014). Other studies have focused on the application of user metrics such as the number of followers, mentions, and retweets (Cha et al., 2010; Pal & Counts, 2011). This identification seems a critical task when managing campaigns on Twitter aiming to have an impact on the offline world. It allows campaigners to identify the most influential users in the online community, which can support the design of strategies aiming either to make visible the supportive users or harm the credibility of opponents (Ceron & d’Adda, 2016; Segesten & Bossetta, 2017).

In the pursuit of becoming influential users, the last decade has seen politicians’ practices of Twitter-use gain sophistication and become more integrated into the overall campaigning approach. Until around 2010, politicians used Twitter basically to broadcast messages, which resembled the traditional unidirectional form

¹ <https://en.oxforddictionaries.com/definition/influence>

of communication while failing to capitalise on the ability to interact with voters (Grant, Moon, & Busby Grant, 2010; Small, 2010). Nowadays, politicians use Twitter to communicate and engage in deep and open interaction with users from different backgrounds (Ausserhofer & Maireder, 2013; Enli & Skogerbø, 2013; Larsson & Ihlen, 2015; Lee & Xu, 2018). This way of communication makes up a rich data source to detect support or opposition from the public. Accordingly, politicians are more eager to engage with users in different ways such as following back new followers (Parmelee & Bichard, 2012), retweeting, replying, and liking tweets from others (Anderson, 2017; Larsson & Ihlen, 2015). Therefore, maintaining and engaging with users is crucial when aiming to measure user sentiment.

The novelty of this study is twofold. First, it introduces a new approach for data pre-processing for sentiment analysis purposes. This approach integrates features of tweets that have been discarded in previous works, namely hashtags, emoticons, and URLs. These features carry information that can help determine the sentiment of tweets, while integrating them in the sentiment analysis can reduce analysis processing time and avoid the occurrence of system crash due to the reduction of data noisiness. System crash refers to the sudden stop of the software used for sentiment analysis purposes. Second, this study responds to an overlooked need to identify influential users from the evidence of the data instead of intuitively. This study provides new evidence that Twitter content offers a convenient approach to measure electoral outcomes and a useful information for identifying actors who set the political agenda.

To test both approaches to data pre-processing and influential user identification on Twitter, this study extracted tweets generated during the 2017 Ecuadorian Presidential election in daily intervals. The study analyses about 1.3 million tweets in Spanish language that 140,617 users posted during the 16 week-period of official campaigning, which were related to the two most voted candidates, Lenin Moreno and Guillermo Lasso. The results of the sentiment analysis are used as indication of vote share during this electoral process and then compared against both the official results and the vote intention prediction by polling firms authorised by the electoral system authority of Ecuador (CNE). This study also identifies the most influential users that were relevant to the candidates every week during the campaign.

The rest of this paper is structured as follows: Section 5.2 introduces the Twitter background and its features. Section 5.3 examines previous works about the

application of sentiment analysis tools and the identification of influential users. Section 5.4 presents the methodology, providing details for the processes of sampling, pre-processing, and analysis of the datasets, followed by the application of sentiment analysis and identification of influential users' approaches. Section 5.5 shows the results and discussion of this work. Finally, Section 5.6 presents the concluding remarks.

5.2. The Twitter environment

Twitter, a microblogging social network founded in 2006, is one of the most popular social media platforms that allow users to generate and spread information in real-time in an effortless way. Twitter allows users to generate posts, which are usually referred to as tweets, within a limit of 280 characters. Besides text, tweets can also contain up to four static pictures, one animated image as a GIF (graphic interchange format) file, a recorded video of no more than 140 seconds², links to external websites, geo-referenced location, and interactive polls. The network is built by adding followers, so a user has both followers and followees. Unlike other social media accounts, Twitter relationship is not automatically reciprocal, meaning that users are not necessarily following each other. Nonetheless, following back can be increased by generating interesting and meaningful tweets to engage with existent and new Twitter user accounts. Lastly, by default, tweets are publicly visible, which means that tweets are also available for non-follower users to see and retrieve. Yet, users have the option to protect their tweets, in which only their followers can have access to them.

Twitter provides different interactive features that allow users to engage and connect with others. These features include retweet, like, reply to tweets, direct messages (DM), mentions, and the use of hashtags, as shown in Figure 5.1. The study of these features is critical for understanding Twitter user behaviours and the content users find interesting, and for pursuing Twitter audiences to spread a message or an intended perception (Meier, Elswailer, & Wilson, 2014; Tang, Ni, Xiong, & Zhu, 2015). Retweet refers to the act of distributing or disseminating tweets within a network. Retweets can be of two kinds. One involves retweeting the original tweet as it is. This act is often understood, among the other users within a network, as a supporting position or endorsement by the user that retweets (Boyd, Golder, & Lotan,

² Paying Twitter subscribers can upload videos up to 10 minutes long.

2010; Garimella, Morales, Gionis, & Mathioudakis, 2018; Metaxas et al., 2015). The second kind involves the addition of a personal comment about the tweet being retweeted and is also known as quote retweet. For the latter, the retweet does not necessarily mean support, since it could include an opposite viewpoint concerning the original tweet.



Figure 5.1. Identification of the interactive features that can be found in tweets.
Translated tweet: "Today we signed the Agreement #PorLaUnidadYElFuturo with @IvanEspinelM. Welcome comrades @Fuerza2017 to this democratic space."

When retrieving tweets for analysis by using the Twitter Search API (application programming interface) to external resources, such as spreadsheet documents, only retweets of the first kind, as introduced in the previous paragraph, are considered retweets. Quote retweets, on the other hand, display the comment followed by an URL (uniform resource locator) redirecting to the original tweet.

The like feature is also known as favourite. It can be used variously to express appreciation and support to the tweet or its author, to show acknowledgment, or as a

bookmark tool to access tweets in a straightforward way through the *Likes* tab in the user profile. This feature invokes an emotional stimulus that users might relate to (Meier et al., 2014). However, it would seem that there are similarities in the motives behind retweeting and liking behaviours, so some users use these features indistinctively. While, to the best of our knowledge no previous research has focused on this, the way features are used might depend on users' personal preferences based on what they value the most. This means that it is not definitive whether liking or retweeting represents a higher scale of support. However, one difference lies in the level of diffusion and distribution of content to reach a wider audience, which is higher when retweeting since retweets are always visible to followers' timelines.

The mention feature refers to the process of tagging another Twitter user account using the handle symbol "@" followed by the username, and can be used for public replying purposes, for promotion purposes to attract more users, and to encourage active interactions when posting tweets (Tang et al., 2015). The mention feature is also used when replying to another Twitter user. As for DM, this feature refers to sending private messages that can be only read and replied by the involved users, and is generally used when sharing sensitive information (He, Lee, & Rui, 2019).

Hashtags constitute another popular feature for gaining more exposure among Twitter users. A hashtag is a keyword which is preceded by the hash sign "#" and can be sometimes a "made-up word" or a composite one. It is commonly used for identifying and categorising topics about specific products, events, users, or places. This means that tweets including hashtags can be easily discovered, which seems meaningful for organisations and users aiming to gain more notoriety and attention from others. When clicking on any hashtag, the most recent tweets that include that hashtag are shown. There are some recommendations about including hashtags in tweets, such as not adding more than two hashtags per tweet and capitalising the first letter of each word to make the hashtag easier to read. For example, using *#PresidentialElection*³ instead of *#presidentialelection*.

Popularity of Twitter comes for different reasons. For example, Twitter allows users to easily generate and consume real-time information, which is useful to engage with potentially difficult-to-reach populations. It provides participants with transparency, anonymity, and a more accessible method to reach widespread

³ <https://business.twitter.com/en/blog/how-to-create-and-use-hashtags.html>

audiences. Popularity also comes from its penetration in the traditional mainstream media coverage, where tweets are often highlighted as source of news. Twitter is a channel that allows users to have open discussions about issues of public interest, such as elections, disasters, or social events (Burgess & Bruns, 2012; Hong et al., 2013), and to facilitate the dissemination of information (Bakshy et al., 2011). From the perspective of business, Twitter's attractiveness is related to the opportunity to advertise products and services by reaching a wider and more global audience, while also increasing brands awareness. Companies also use Twitter to engage with other users to understand how brand perception can influence customer feelings, which can be reflected in the maximisation of profits. In addition, its cost-effective means of data sampling, and the straightforward way to process and analyse data are relevant when it is aimed to design campaigns with an impact on both online and offline worlds.

5.3. Related work

Arguably, Twitter is to date the most convenient social media platform to develop sentiment analysis models in the political arena. This is despite Instagram being the fastest growing social media platform (Na & Kim, 2019) and Facebook remaining as the most popular (Kachamas, Akkaradamrongrat, Sinthupinyo, & Chandrachai, 2019). Due to the short length of tweets, users have limited room to justify their viewpoints, which forces them to be as precise as possible. Therefore, Twitter discourages vagueness of meaning when producing content. As a result, conversations can be more heated than on other social media platforms, especially for the political arena (Auvinen, 2012; Conover et al., 2011a).

Twitter has been used by academics and campaigners as a source to develop models based on sentiment analysis. Sentiment analysis usually incorporates the use of natural language processing, statistics, and text analysis (Sweeney & Padmanabhan, 2017), and it has been proved to be an effective tool for summarising and coding opinions posted on Twitter (De Clercq & Hoste, 2016). In this regard, the use of sentiment analysis seems relevant for understanding insights into user behaviours. Models relying on sentiment analysis using Twitter data have been broadly used, as described in the following section.

By leveraging sentiment analysis models from Twitter data, consistent results have been achieved in different fields. In the commercial arena, for instance,

sentiment analysis has been used to estimate box-office revenues for movies by considering the rate and sentiment of tweets (Asur & Huberman, 2010; Jain, 2013). Alternatively, the number of followers and positive tweets have been associated with higher movie sales (Rui, Liu, & Whinston, 2013). Concerning stock markets oscillations, studies have found that public mood correlates with the market index (Bollen et al., 2011; Fanzilli, 2015; Smailović et al., 2013). In the health field, sentiment analysis tools have been applied to study mental health (Coppersmith, Dredze, & Harman, 2014), smoking behaviour (Myslin, Zhu, Chapman, & Conway, 2013), and to detect the incidence rate of influenza by applying overall tweet count (Aramaki et al., 2011; Paul, Dredze, & Broniatowski, 2014; Santos & Matos, 2014). In the sports arena, it has been used to determine match outcomes in the English Premier League, which has led to higher pay-out returns in bets, although with lower accuracy (Schumaker, Jarmoszko, & Labeledz Jr, 2016).

In the political arena, there are different approaches to sentiment analysis that can help detect the vote choice. With the assistance of corpora, sentiment analysis outcomes have been achieved by studying the overall polarity of tweets (Burnap et al., 2016), by counting tweets mentioning a political candidate or party as result of voter outcome (Borondo, Morales, Losada, & Benito, 2012; Caldarelli et al., 2014; Tumasjan et al., 2010), and by tweet count but removing duplicated tweets or users (Sang & Bos, 2012). Also, calculating the size of online networks has been used to estimate election outcomes (Cameron, Barrett, & Stewardson, 2016). On the other hand, analysis of political preferences has been conducted by examining the structure of networks of political retweets to classify a user's political alignment (Barberá, 2015; Conover et al., 2011b). Both machine learning approaches (Anjaria & Guddeti, 2014; Cameron et al., 2016; Smailović, 2014) and lexicon-based methods have been applied to develop sentiment analysis tools to estimate election results (Jose & Chooralil, 2015; Parackal, Mather, & Holdsworth, 2018; Ramteke et al., 2016; Tsakalidis, Papadopoulos, Cristea, & Kompatsiaris, 2015).

5.3.1. Building models through sentiment analysis tools

As Twitter users continuously produce and consume information, suppliers and merchants observe and record patterns of their behaviours to attempt improved satisfaction delivery and return maximisation (Hoang & Mothe, 2017). To do this, one of the functionalities that Twitter can provide is as a foundation for the development

of sentiment models aiming to measure and anticipate user behaviour. Generally speaking, the purpose of sentiment modelling involves the identification of trends for the possibility to intervene in variables to increase the probability of a desired outcome (Schumaker, Solieman, & Chen, 2010a). This is often achieved through probabilistic simulations of performance, which basically rely on mined historical data analysed through knowledge management techniques (Schumaker, Solieman, & Chen, 2010b).

Models of sentiment analysis demand the development of algorithms that can classify sentiment. They are often referred to in the literature as classifiers. Sentiment classifiers are relevant in the political arena to estimate the level of support or rejection towards candidates. Research has focused on incorporating different features and approaches to improve the ability of the classifier to categorise or detect the feeling of the text, and hence, the overall sentiment analysis results.

Before developing the sentiment classifier, it is necessary to pre-process the data. This practice is also known as *cleaning* process. It consists of transforming the less readable content into readable for the algorithms. Some approaches used for pre-processing data include text tokenisation, part-of-speech (POS) tags, bag-of-words (BOW) methods, stemming, or n-grams construction.

The ability of the sentiment classifier to perform well depends on the quality of input data used for its development (Barbosa & Feng, 2010; Gokulakrishnan et al., 2012; Haddi et al., 2013). Pre-processing approaches are of paramount importance to develop reliable classifiers and sentiment analysis models. Pre-processing Twitter data also involves the treatment of hashtags, emoticons, and URLs (Barbosa & Feng, 2010; Saif, He, & Alani, 2012). In some cases, these features have not been considered when sentiment analysis is conducted since it is believed that they do not add much value to the sentiment analysis, and instead, increase the noisiness of the data (Al Hamoud et al., 2018; Jain & Jain, 2019; Pak & Paroubek, 2010; Pandarachalil et al., 2015). Noisiness refers to the data that do not provide value when conducting sentiment analysis as previously introduced in Section 5.1. In other studies, they have been either substituted by relying on existing manually classified emoticons and hashtags datasets to determine the sentiment (Agarwal et al., 2011; Go, Bhayani, & Huang, 2009a; Kouloumpis, Wilson, & Moore, 2011), or replaced by the words that represent them, especially for the URL feature through POS techniques (Saif et al., 2012). However, discarding these features might affect the detection of the

sentiment of the content, since chances are that they hold rich information and summarise the overall feeling towards the subject. A summary of studies applying a pre-processing stage for sentiment analysis predictions is shown in Table 5.1.

Table 5.1

Related work about Twitter feature pre-processing when performing sentiment analysis

Authors	Twitter features				
	Mention	Emoticon	Hashtag	URL	Number
Go et al. (2009)	Replaced by the token (USERNAME)	Replaced by the sentiment polarity	Not specified	Replaced by the token (URL)	Not specified
Pak and Paroubek (2010)	Removed	Replaced by the identified sentiment polarity and then removed	Not specified	Removed	Not specified
Agarwal et al. (2011)	Replaced by the tag T	Replaced by the sentiment polarity using a manually labelled dictionary using the emoticons listed on Wikipedia ⁴	Count of number of hashtags	Replaced by the tag U	Not specified
Kouloumpis et al. (2011)	Replaced by a token	Replaced by the sentiment polarity using the Stanford Twitter sentiment corpus from Go et al. (2009), and then replaced by a token	Replaced by the sentiment polarity using the Edinburgh Twitter corpus ⁵ , and then replaced by a token	Replaced by a token	Not specified
Saif et al. (2012)	Not specified	Replaced by the sentiment polarity using the Stanford Twitter sentiment corpus from Go et al. (2009)	Not specified	Replaced by the token URL	Not specified
Gokulakrishnan et al. (2012)	Replaced by the class <USER>	Replaced by either “happy” or “frown” keywords	Replaced by the class <HASHTAG>	Replaced by the class <URL>	Not specified

⁴ https://en.wikipedia.org/wiki/List_of_emoticons

⁵ <http://demeter.inf.ed.ac.uk>

Authors	Twitter features				
	Mention	Emoticon	Hashtag	URL	Number
Khan et al. (2014)	Removed	Replaced by the sentiment polarity using the Enhanced Emoticon Classifier	Removed	Removed	Not specified
Dwi Prasetyo and Hauff (2015)	Removed	Removed	Not specified	Removed	Not specified
Pandarachalil et al. (2015)	Removed	Replaced by the sentiment polarity using the sentiment tokeniser developed by Christopher Potts ⁶	Removed	Removed	Not specified
Ramteke et al. (2016)	Removed	Not specified	Not specified	Removed	Not specified
Parveen and Pandey (2016)	Removed	Replaced by the sentiment polarity	Removed	Removed	Not specified
Kharde and Sonawane (2016)	Removed	Replaced by the sentiment polarity	Removed	Removed	Removed
Sharma and Moh (2016)	Removed	Removed	Removed	Removed	Not specified
Tunggawan and Soelistio (2016)	Preserved without modification	Not specified	Preserved without modification	Removed	Not specified
Zhao and Gui (2017)	Not specified	Not specified	Not specified	Removed	Removed
Al Hamoud et al. (2018)	Removed	Not specified	Removed	Removed	Not specified
Budiharto and Meiliana (2018)	Removed	Not specified	Not specified	Removed	Removed
Kumar and Babu (2019)	Not specified	Removed	Removed	Removed	Not specified
Nazir et al. (2019)	Not specified	Not specified	Count of number of hashtags	Removed	Not specified
Jain and Jain (2019)	Removed	Not specified	Not specified	Removed	Removed

⁶ <http://sentiment.christopherpotts.net/>

5.3.1.1. *Methods for conducting sentiment analysis*

After the pre-processing stage, the next step is developing and testing the sentiment analysis tools. Research using Twitter data for sentiment analysis purposes has mainly relied on two approaches: supervised machine learning and lexicon-based methods (Neethu & Rajasree, 2013; Saif et al., 2012; Saif, He, Fernandez, & Alani, 2016). The first technique refers to the application of supervised machine learning (language-based) for classification purposes through the use of machine learning approaches, such as support vector machine (Neethu & Rajasree, 2013; Sharma & Dey, 2012; Smailović et al., 2013), Naïve Bayes (Gautam & Yadav, 2014; Pak & Paroubek, 2010; Skuza & Romanowski, 2015), or maximum entropy classifier (Hutto & Gilbert, 2014). Sentiment supervised machine learning refers to the categorisation of the polarisation of tweets using a two-point scale, positive and negative, or including neutral as the third sentiment. After the data are extracted, this method relies on large amounts of manually annotated data, for which the data need to be split into training and testing datasets. In the training dataset, the machine learning method is used to learn patterns, which are tested in the unseen data in the testing subset (Mukhtar, Khan, & Chiragh, 2018).

The second technique is based on lexicon-based methods (knowledge-based), also known as the unsupervised approach, which allows the identification of sentiment polarisation of text through dictionaries of opinion words or corpora, with predefined weights or scores to determine the sentiment polarity of texts (Saif et al., 2016; Sweeney & Padmanabhan, 2017; Taboada et al., 2011). Examples of available corpora are *SentiWordNet* (Baccianella et al., 2010), or *SentiStrength* focused on social media data (Thelwall, Buckley, & Paltoglou, 2012; Thelwall et al., 2010). Lexicon-based approaches for identifying sentiment involve a twofold level of analysis: entity-level and tweet-level (Saif et al., 2016). Entity-level is the identification of feelings considering the mention of users, organisations, or events. For example, “*I don’t usually love ice cream, but Gino is so good!*”. The polarisation starts being negative, but then it turns positive towards the brand Gino. Tweet-level refers to the identification of sentiment based on individual tweets. Evidence shows that when working with Twitter data, the lexicon-based approach works better than the supervised machine learning approach (Choi & Lee, 2017; Mukhtar et al., 2018). Models grounded on lexicon-based techniques do not necessarily require data pre-processing (Dhaoui, Webster, & Tan, 2017; Saif et al., 2016), but it is recommended

to implement it since it can improve the overall performance of the sentiment analysis models, while reducing computational processing time.

A critical component when conducting sentiment analysis is the availability of corpora or labelled dictionaries to develop the models, which is not frequent for non-English languages. A corpus in the sentiment analysis refers to the collection of written structure material that has identified sentiment polarisation associated with the text. Even when corpora are available in a language different from English, chances are that they might not be complete enough to perform well, especially in the political context (Jha, Manjunath, Shenoy, & Venugopal, 2016). In addition, since a corpus is composed mainly of formal words, there might be the use of slang and local language, which may not be known or integrated in the corpus, thus affecting sentiment classification. Some studies have created their own dictionaries by collecting tweets and performing manual detection of sentiment, either by analysing texts or just the emoticons found in tweets (Neethu & Rajasree, 2013; Pak & Paroubek, 2010), but this can be time-consuming and label-intensive (Liu & Zhang, 2012; Mukhtar et al., 2018), requiring also the participation of different labellers to obtain more accurate sentiment. To overcome this limitation, the use of automated sentiment-detection software can be an option when corpora are not fully available.

5.3.1.2. Performance evaluation of sentiment analysis with real cases

When comparing the performance of Twitter data with that of traditional polls, with reference to sentiment analysis in the political context, different conclusions have been reached. A number of academics have argued that Twitter data are more convenient because provide more accurate, cost efficient, and timely results than traditional polls (Ceron, Curini, & Iacus, 2015; Godin et al., 2014; Sang & Bos, 2012). Others have stated that Twitter data highly correlate with polling results, so Twitter can be seen as a useful complement to offline polls (Borondo et al., 2012; O'Connor, Balasubramanyan, Routledge, & Smith, 2010). Finally, there are those that reject the convenience of Twitter by arguing instead that sentiment in tweets is not as sharply defined as it is in traditional polls, nor it is indicative of poll results (Bermingham & Smeaton, 2011a; Mejova, Srinivasan, & Boynton, 2013; Mellon & Prosser, 2017; Mitchell & Hitlin, 2013; Schoen et al., 2013; Sinnenberg et al., 2017). However, a key limitation of the survey approach is that it involves high operational costs and estimations of the sentiment are not dynamically updated (Jin et al., 2010),

which places a burden on the ability of campaigners to implement opportune and effective action.

In addition, some critiques that have been made against Twitter data for sentiment detection purposes include the inability to identify fake accounts, little attention to demographic variables, the unknown effects of sampling approaches, and the issue of self-selection bias (Gayo-Avello, 2012; Jungherr, Jürgens, & Schoen, 2012; Metaxas, Mustafaraj, & Gayo-Avello, 2011). Furthermore, the fact that the sentiment of words often depends on the syntactical and semantic language rules (Cambria, 2013) and on the context in which they are used (Liu & Zhang, 2012), adds further criticism to the use of Twitter data as source for identifying users' feelings. Similarly, the sentiment of an irony or sarcasm, slang, abbreviation, and misspelling are hardly detected, which can affect the ability of a model to perform well (Neethu & Rajasree, 2013). However, during the last few years, the power of Twitter for detecting the sentiment during electoral campaigns has been re-evaluated, confirming that, despite the challenges mentioned above, it provides a valuable and effective source of data that can beat that of traditional polling (Cody, Reagan, Dodds, & Danforth, 2016; Le et al., 2017; Wang & Gan, 2018).

Finally, not all sentiment detection models have been accurate in anticipating election's results. For example, Burnap et al. (2016) were not able to anticipate, by means of Twitter, the party that would win the majority of seats in the Parliament before the 2015 General election in United Kingdom. However, polling companies also failed to deliver accurate results, even when they used different techniques such as face-to-face surveys and telephone polls. From the viewpoint of Hodges (2015), results can go wrong because of manipulation of the information and perceptions by mainstream media; but Clark (2015) suggested that these outcomes might not be accurate because pollsters can fail to properly address sampling factors and demographic aspects, and Healy (2015) claimed that the strategies used by the polling companies can be simply not adequate. Thus, sentiment detection models are not infallible methods, because there might be neglected factors that could affect the overall outcome, or just because in politics people can be persuaded to change their minds in the last minute for unforeseen reasons.

5.3.2. Who is setting the agenda in political aspects?

The identification of actors that are setting the political agenda of Twitter is

essential for understanding how the political system and diffusion of information during elections work (Aral & Walker, 2012). Before the use of the Internet, traditional mainstream media organisations were the entities mainly responsible for the propagation of information as well as the setting of political agendas (McCombs, 2014; Owen, 2017). It is assumed that they have the power to shape users' behaviours and capture user's attention, influencing to some extent decision-making processes (Halper, 2016).

To adapt to technological changes in the marketplace, mainstream media have been delivering content through the Internet. Similarly, politicians and campaigners have turned to social media content because they are aware that this is, perhaps, the most powerful mechanism to disseminate information. Moreover, it is likely that politicians prefer to manage the information that their like-minded users consume without journalists' involvement (Shafi & Vultee, 2016). In addition, social media platforms are the venue where users can generate and discuss political content, influencing the ability to connect with others politically.

With the increasing use of Twitter as a platform to produce, consume, and disseminate political content, a number of users have emerged, with the power to spread content fast throughout their networks and beyond. On Twitter, these users are regarded as influential users. Alp and Ögüdücü (2018), and Aral and Walker (2014) agree that influential users have the ability to persuade others' behaviours by the use of recommendations, spread of information, or viral marketing techniques. Similarly, influential users have the power to set the political agenda, in part because other users see them as experts on related topics. Moreover, in the political arena, the evidence suggests that the most influential users on Twitter are politicians, established journalists, media firms, and bloggers (Dubois & Gaffney, 2014; Larsson & Moe, 2012; Parmelee, 2014). Therefore, Twitter contributes to the diffusion of political debates (Larsson & Moe, 2012), and provides a valuable platform for users to discuss public political issues and uncover content that is overlooked by traditional media and politicians.

The debate about what constitutes an influential user on Twitter is far from closed. Different approaches are used to conclude whether a user is influential or not. To mention two of them, social network analysis (SNA) tools are one of the most used approach to identify influential users within networks (Cossu et al., 2015; Kitsak et al., 2010; Smith, Rainie, Shneiderman, & Himelboim, 2014; Xu et al., 2012), which

can be either visually represented through graphs or matrices. The SNA approach involves the identification of online networks that allow the discovery of patterns and connections in social relationships, providing an understanding of their structural properties (Scott, 2017). Besides SNA and graph techniques, there have been different ways to categorise a user as influential. For example, some studies have relied on the concept of followership, defined by the number of followers of a Twitter user account, as the main measure of influence, since it is believed that users with larger number of followers have more influence than other users (Bae & Lee, 2012; Kwak et al., 2010; Weng, Lim, Jiang, & He, 2010). Other studies have used the number of retweet and mentions as measures of influence (Cha et al., 2010; Linh et al., 2013). Retweet influence is associated with those users that have the power or ability to spread messages widely (Rattananitnont, Toyoda, & Kitsuregawa, 2012; Rogers, 2010). On the other hand, *mention* influence is an indication of the ability of a user to engage others in a conversation, and embodies the name value of Twitter users (Cha et al., 2010; Tumasjan et al., 2010). Hence, there is no single answer to the question of what constitutes an influential user on Twitter. Yet, these different attempts to define influential users have shed light on the possible roles played by users within their networks.

Besides the identification of influential users, the identification of the most frequent types of tweets during political campaigns is likewise important, because it provides a glimpse of how users interact with others. This task is known as identification of the network distribution. A number of studies have focused on the types of tweets, which have been categorised as singleton or normal post, reply, retweet, retweet with comment, and mention (Bruns, 2012; Graham, Broersma, Hazelhoff, & Van'T Haar, 2013; Larsson & Moe, 2012; Russell, Hendricks, Choi, & Stephens, 2015; Tumasjan et al., 2010). These studies have focused on visually mapping dynamic conversation to understand the use of Twitter during political events, and to create topologies about how politicians and users interact on Twitter. Finally, the identification of both influential users and network distribution, has been conducted as two independent analyses. Also, the sentiment associated with the most influential actors within networks has received little attention in the literature. These tasks are all important for a deeper understanding of user behaviours during political campaigns.

5.3.3. Contribution of this paper

The purpose of this study is twofold. First, it contributes to the myriad of emerging pre-processing approaches aimed at enhancing the potential of Twitter data for modelling sentiment classifiers. Since the messiness of Twitter content continues to challenge the accuracy of Twitter-based models that rely on sentiment analysis, especially for political purposes, this study presents a novel data pre-processing approach considering features embedded in tweets which have been overlooked in previous works. The proposed approach improves the performance of sentiment analysis, reduces time and operational costs, and avoids system crashes. Second, it contributes to unpacking the dynamics of the identification of influential users on Twitter that may influence the electoral preferences of users. For this reason and using the same dataset as for the sentiment analysis model, a novel approach for identifying influential users is presented, which detects the most frequent type of tweets, and from this, it identifies the most relevant and influential Twitter user accounts setting the agenda in political debates, and their sentiments towards each candidate. The information produced herein can be useful for managing Twitter campaigns by enhancing the ability of sentiment analysis tools to detect the sentiment on Twitter, and by periodically identifying influential users during electoral campaigns setting the political agenda.

5.4. Methodology

This section first presents the case study and describes the approaches used for data sampling, pre-processing, and analysis, which constitute the basis for the sentiment analysis task. Then, it presents a novel approach to periodically identify the most influential Twitter users setting the political agenda relevant to each of the candidates.

5.4.1. Case study: The 2017 Presidential election in Ecuador

This study uses Twitter content relevant to the 2017 Ecuadorian Presidential election produced during the official campaigning period. While there is no agreement in the literature on what constitutes a good case for studying social media content in relation to data and decision science development, this data source is of particular value to the purpose of this study for two reasons. First, the Presidential elections in Ecuador are one of the topics producing great amounts of social media

content. The previous president, Rafael Correa, was the most influential Twitter user in Ecuador in terms of number of followers, while the official presidential account on Twitter was the most popular in the world in this category, adjusted for population (Carvajal, 2017; González, 2016; Miceli, 2015; Socialbakers, 2017). And second, user deliberations about different dimensions of the candidates were particularly intense across the tight presidential race, in which social media fuelled a context of dirty campaigning that ended up in a sort of political hysteria (González, 2017; Stoessel, 2017). In addition, conducting this study in the heart of Latin America contributes to cultural and language contexts still little explored in research relevant to decision and data science.

5.4.2. General information about Ecuadorian elections

Presidential elections in Ecuador use a two-round system. During the first round a candidate can become president under two scenarios: (1) obtaining more than 50% of the votes, or (2) getting over 40% of the vote and being 10% ahead of the nearest contender. Otherwise, a second round is conducted with the two most voted candidates, where the winner is the one with the highest number of votes. Voting is mandatory for eligible voters from those older than 18 up to 65 years old. There is a constitutional exemption rule for those between 16 and 18 years old, those older than 65, Ecuadorian citizens living outside the country, active duty army and police members, handicapped, and illiterate people. This exemption also covers foreign people from 16 years old, legally residing in the country at least for 5 years, and enrolled in the electoral register.

By 2017, Ecuador had a population of sixteen million people, with nearly 12.4 million registered voters. The voters' age distribution during the second round of the presidential election 2017 shows that more than 80% of voters were from the compulsory age range⁷ (see Figure 5.2). During the first round, eight candidates vied to succeed president Rafael Correa. The two most voted candidates were Lenin Moreno from the ruling party, and Guillermo Lasso from the opposition. Having finished the first round, and without any of the candidates meeting the threshold to

⁷ <https://app03.cne.gob.ec/EstadisticaCNE/Ambito/Distributivo/Distributivo.aspx>

avoid a run-off, a second round was required one month later with the two candidates, where Lenin Moreno was elected president with 51.15% of the vote⁸.

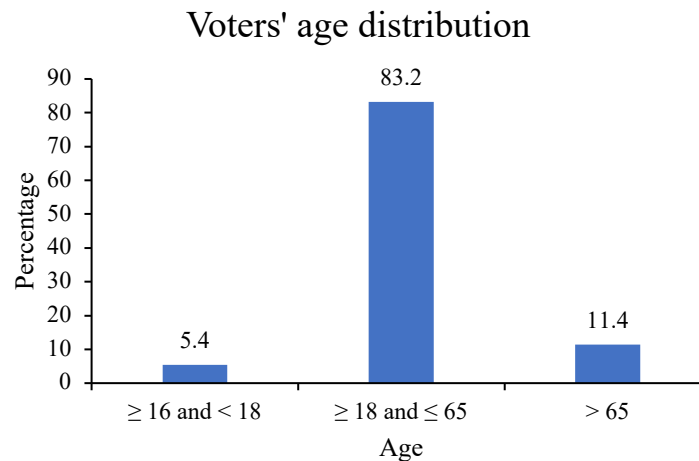


Figure 5.2. Electorate's age distribution during the second round of elections, 2017.

5.4.3. Data sampling

This study uses 1,356,728 tweets in Spanish language, relevant to the 2017 Ecuadorian Presidential election, posted by 140,617 users during the official campaigning period. By comparison, the polling firms had used samples of between 2,000 and 3,000 voters to estimate the vote share of the 12.4 million registered voters. The extracted tweets were publicly available, meaning that those Twitter user accounts which marked their tweets as private could not be accessed or retrieved. The data extraction for the first round took place between the 26th of November 2016 and 17th of February 2017, and for the second round was between the 5th of March and 1st of April 2017, comprising 16 weeks in total. These dates were selected because they were consistent with the official starting and ending dates for each of the two rounds. Moreover, there is a one-month gap of data extraction between the two rounds, because during this period, the electoral system authority of Ecuador (CNE) announced the results of the first round and campaigning was not authorised.

The criteria for data sampling were to retrieve those tweets that included information about the two favourite candidates, Lenin Moreno and Guillermo Lasso, by using mentions (@*Lenin*, @*LassoGuillermo*), hashtags (#*leninmoreno*, #*guillermolasso*), and keywords including candidates' names. There are other words

⁸ <http://cne.gob.ec/es/institucion/sala-de-prensa/noticias/3994-pleno-del-cne-presento-resultados-totales-de-la-segunda-vuelta-electoral>

with which the candidates are associated, comprising in most cases derogatory or offensive names, which were not considered for analysis. Such words can add some noisiness to the model and thereby affect the ability to detect the sentiment. In the data analytics arena, noisiness refers to the inclusion of content that can add sensitivity to error (Kumar et al., 2014). In the context of this study, the offensive names referred to above are likewise used in contexts far different from that of the 2017 Ecuadorian Presidential election. Hence, using those names as keywords for retrieving tweets could lead to the retrieval of irrelevant content.

For the extraction of tweets, the study relied on the Twitter Search API tool and a collection of R scripts (R Core Team, 2013). Tweets were extracted on a daily basis. However, since the Twitter Search API tool has a restriction of extracting only the 3,200 most recent tweets every time, an R function was implemented to run the extraction script every three hours throughout the campaign. Then, the data were classified in weekly intervals for analysis. In addition, the numbers of followers and followees for both candidates were collected at the end of every week to visualise their evolution throughout the campaign. Lastly, all tweets retrieved were assumed as having come from a trustworthy source. This means that this study assumed Twitter users as truthful and as being actual individuals involved in the debates relevant to the 2017 Ecuadorian Presidential election. Hence, the study did not consider the presence of fake users or bots, nor actual users producing fake news. This is a limitation in that there is some evidence that suggests the existence of these types of accounts during this election (Rofrío et al., 2019). The effect of the latter types of users on the sentiment detection task should be addressed in future research.

An example of the appearance of a tweet from the Twitter interface is shown in Figure 5.3.



Figure 5.3. Appearance of a tweet from Guillermo Lasso. Translated tweet: “*I do not intend to be the rector of morals or the spiritual director of anyone. I aspire to be the President of Ecuador #DialogoUSFQ*”.

After the tweet was retrieved using R, it was transformed into a dataframe containing the attributes as shown in Table 5.2.

Table 5.2

Appearance of the tweet in its original format after retrieving it through the Twitter Search API

text	Yo no pienso ser el rector de la moral ni director espiritual de nadie. Yo aspiro a ser Presidente del Ecuadorâ€ https://t.co/h2Ht77AOoF
favorited	FALSE
favoriteCount	385
replyToSN	NA
created	23/03/2017 13:08:00
truncated	TRUE
replyToSID	NA
id	844974000000000000.00
replyToUID	NA
statusSource	Twitter for iPhone
screenName	LassoGuillermo
retweetCount	466
isRetweet	FALSE
retweeted	FALSE
longitude	NA
latitude	NA

text: Text of the tweet.

favorited: Boolean value. True if the tweet has been marked as favoured; otherwise, false. This attribute appears to be deprecated.

favoriteCount: Number of times the tweet has been liked by Twitter users.

replyToSN: If the tweet is a reply, it will show the username of the original tweet's author; otherwise, NA (not applicable).

created: Date and time the tweet is created.

truncated: Boolean value. True if the text of the tweet has been truncated when retrieved; otherwise, false.

replyToSID: If the tweet is a reply, it contains the numerical identification for original tweet; otherwise, NA.

id: Unique numerical identification for the tweet.

replyToUID: If the tweet is a reply, it contains the numerical identification for the original account; otherwise, NA.

statusSource: Source where the tweet was created. It can be the name of the device, such as Android, or website addresses.

screenName: User's screenname.

retweetCount: Number of times the tweet has been retweeted by other Twitter users.

isRetweet: Boolean value. True if the tweet is a retweet from another tweet; otherwise, false.

retweeted: Boolean value. True if the tweet has been retweeted; otherwise, false. This attribute appears to be deprecated and it is not used in this study.

longitude: If geolocation is enabled, it will show the longitudinal coordinate where the tweet is created; otherwise, NA.

latitude: If geolocation is enabled, it will show the latitudinal coordinate where the tweet is created; otherwise, NA.

5.4.4. Data pre-processing

Data pre-processing is a task aimed to enhance the accuracy of the sentiment analysis (Haddi et al., 2013). To assess this construct, this study compares the sentiment of tweets detected before and after pre-processing the data, against the official results of the 2017 Ecuadorian Presidential election, whereby tweets with positive sentiments towards a candidate are proxy of vote choice. In this study, the data pre-processing involved cleaning and organising the data before performing the

sentiment analysis. Numbers, mentions to Twitter user accounts, punctuation signs, extra blank spaces, and special characters were removed from the tweets. Also removed were the tweets written in a language other than Spanish, which accounted for 583 tweets (0.04%).

The next task involved selecting and coding the other-than-text features of tweets to include in the sentiment analysis. As previously shown in Table 5.1, researchers have often discarded these features before conducting sentiment analysis. When considered, previous studies have treated these as numeric values. For example, Nazir et al. (2019) analysed the frequency with which hashtags appeared in tweets. Discarding these features can be problematic since besides overlooking informational value, the use of hashtags, emoticons, and URLs are a very popular way of communicating in the Twitter environment (Tumasjan et al., 2010; Stoicescu, 2016).

Hashtags, besides indicating the subject of a tweet, constitute an important component of its overall sentiment (Davidov, Tsur, & Rappoport, 2010; Rezapour, Wang, Abdar, & Diesner, 2017). In the political context, furthermore, a hashtag is commonly used to promote a sense of identity, belongingness, and attachment to a party or candidate (Khan, Zaher, & Gao, 2018; Zhou, 2011). Emoticons permit users to add an emotional touch to a message (Amaghlobeli, 2012), which reveals or gives clues of the sentiment of a tweet being positive or negative (Joshi, Simon, & Murumkar, 2018). Concerning URLs, these are widely used on Twitter to share information that cannot be thoroughly communicated by written means, in part, due to the short lengths of tweets (Chy, Ullah, & Aono, 2015). The popularity of URLs is based on the mechanism offered to summarise emotions and thoughts about a subject or a Twitter user account (Cui, Zhang, Liu, & Ma, 2011). In summary, the three features mentioned above, namely hashtags, emoticons, and URLs, constitute data that carry both informational value and sentiment, and therefore should not be ignored or evaded when conducting sentiment analysis.

To pre-process the hashtags, the datasets were first exported into spreadsheet documents in Microsoft Excel. Then, a code was implemented in Microsoft Excel's Visual Basic Editor to split hashtags into independent words. For example, suppose a hashtag “*#ISupportLenin*” was transformed into the text “*I Support Lenin*”. A limitation of this approach is that it performs the splitting task only when the first letter of each word composing the hashtag is uppercase. While this is a common practice on Twitter, however, a minor proportion of the hashtags contained only

lowercase letters and thereby could not be transformed into readable text. The latter kinds of hashtags were then removed from the datasets for analysis.

Concerning the pre-processing of emoticons, previous research has replaced emoticons, such as :) or :(, with their equivalent words, but this is limited to simple representation of emoticons in form of typography to represent facial expressions. Nowadays, Twitter can support emoji, e.g. 😊 or 😞, which are described as pictographs to describe situations, and are based on the Unicode standard for character encoding (Hern, 2015). When tweets containing emoji are extracted from Twitter through a programming language such as R, the picture is transformed into unique codes, which can be in terms of bytes (UTF-8) or R-encoding as shown in Table 5.3. For this study, emoticon also includes the term emoji, and they were transformed into words in Spanish by adapting the decoder file developed by Peterka-Bonetta (2015). After the replacement, the codes were removed from the datasets.

Table 5.3

Example of emoticons and emoji coding

Description	Emoticon	Emoji	Unicode	UTF-8	R-encoding
Slightly smiling face	:)	😊	U+1F642	0xF0 0x9F 0x99 0x82	<ed><U+00A0>< U+00BD>
Slight frowning face	:(😞	U+1F641	0xF0 0x9F 0x99 0x81	<ed><U+00B9>< U+0082><U+263 9><U+FE0F>

Regarding URLs, it was noticed that their frequencies followed a power-law distribution, meaning that few URLs appeared in a large number of tweets while the vast majority appeared only in a few tweets. This observation is shown in Figure 5.4 and Figure 5.5 for each candidate during the 16 weeks of the campaign. The y-axis shows the frequency with which URLs were shared, while the x-axis displays each URL in order from most to least shared. For example, Figure 5.4 shows that approximately 75,000 URLs were shared in tweets relevant to Lenin Moreno during the campaign. However, while the most shared URL appeared in 2,225 tweets, the vast majority of them appeared in only a few tweets. This power-law distribution is observed even when tweets are grouped into weekly intervals throughout the campaign, although the frequencies increase in the last weeks of each round, as shown in Figure A-5.1 and Figure A-5.2 in the Appendix A-5. In these figures, the scales of

the plots differ every week for each candidate because the frequency with which URLs were shared varied. For example, in Figure A-5.2, the most frequent URL in the third week of the campaign for Guillermo Lasso was shared about 150 times, whereas the most frequent URL in the week 15 was shared about 3,000 times. Thanks to this power-law distribution of URL frequencies, we were able to detect the sentiment of 80% of the tweets containing URLs by manually determining the sentiment of the few most frequent URLs. Then, each URL was replaced with the sentiment in the dataset for analysis, being it positive, negative or neutral.

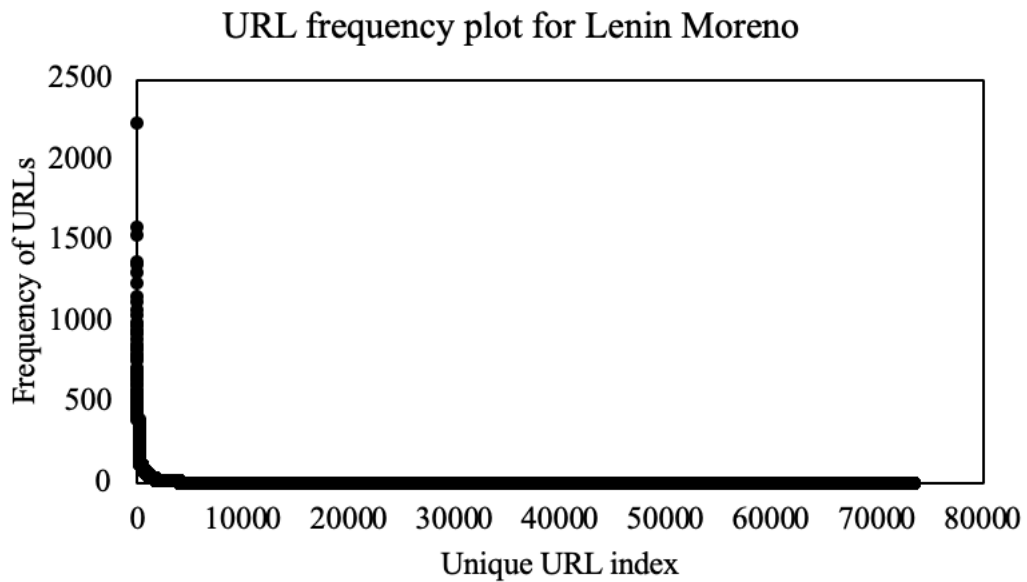


Figure 5.4. Frequency plot of URLs found in tweets about Lenin Moreno.

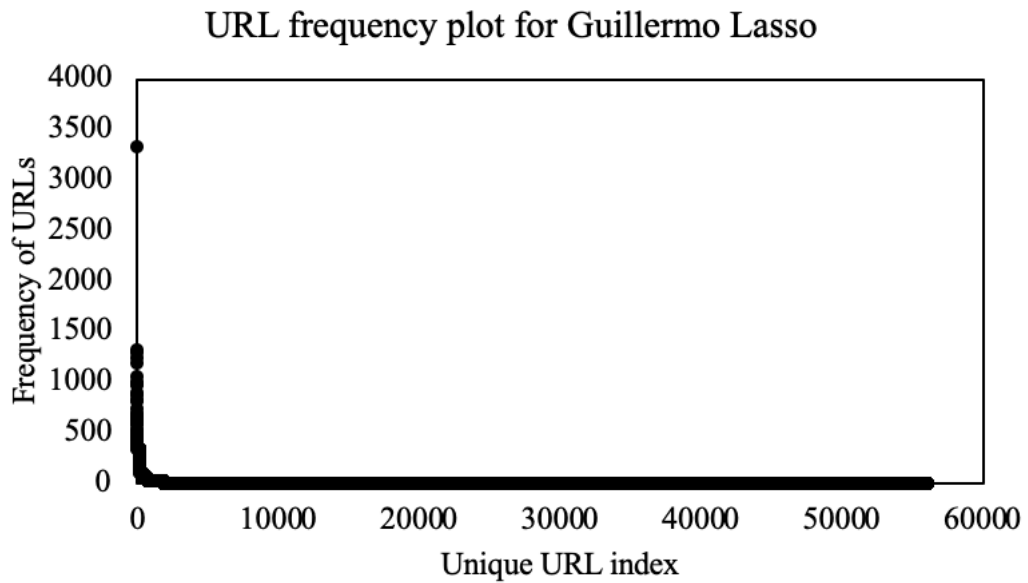


Figure 5.5. Frequency plot of URLs found in tweets about Guillermo Lasso.

5.4.5. Sentiment analysis

To perform the sentiment analysis, several text analytical software tools were evaluated based on their technical capabilities to deal with Spanish language content, including service providers Semantria®, Repustate, and MeaningCloud™. These three lexicon-based software providers were evaluated during a trial period using a sample of 2,000 tweets from the extracted data. To begin with, a sentiment label of positive, neutral, or negative was manually assigned for each tweet of the sample. A positive or negative label means that a tweet shows explicit support or rejection to a candidate, whereas neutral means that no clear position is stated in the tweet. Then, the sentiment detection task for the tweets was performed with each of the providers' applications. As a proxy for accuracy, the proportion of tweets that matched the sentiment label applied manually was recorded for each of the providers. In the end, 88% of the sentiment labels obtained with MeaningCloud™ matched the labels assigned at the beginning, followed by Semantria with 74% and Repustate with 68%. Consequently, MeaningCloud™ was selected as the sentiment analysis provider.

MeaningCloud™ is implemented as an add-in for Microsoft Excel. It uses advanced natural language processing techniques to detect polarity in texts⁹. As a bonus feature, it allows users to personalise their own dictionaries according to the

⁹ <https://www.meaningcloud.com/blog/an-introduction-to-sentiment-analysis-opinion-mining-in-meaningcloud>

context of the study. Accordingly, a personalised dictionary adding local slang to include user-defined concepts for the sentiment categorisation was implemented and integrated for the sentiment analysis. This means that slang and other nonstandard words that are commonly used in everyday language in Ecuador were included, thereby adding new semantic meanings and the corresponding sentiment. Some examples are “*comecheques*” or “*APes*” which are made-up words that emerged during the campaign to convey negative sentiment towards one of the candidates. Lastly, MeaningCloud™ has been used for academic purposes in tasks involving sentiment analysis in Spanish language (Bilro, Loureiro, & Guerreiro, 2018; Loyola-González et al., 2019; Zanfardini, Biasone, Bigné, & Ferri, 2015; Zanfardini, Bigné, & Andreu, 2017).

Among different services, MeaningCloud™ performs multilingual sentiment analysis of texts, and delivers six categories for the polarisation of the tweets, which are very positive, positive, neutral, negative, very negative, and none. A tweet is labelled as neutral if both polarities, positive and negative, are found in the text, while the none sentiment is assigned when MeaningCloud™ cannot detect any polarity. Since this study focuses on detecting if the sentiment of tweets is positive, neutral, or negative to use as a proxy for vote choice, upon running MeaningCloud™, the tweets labelled as very positive were merged with positive, very negative with negative, and none with neutral. This means that tweets labelled as “very positive” and “positive” equally imply support, whereas those as “very negative” and “negative” equally imply opposition. Hence, three sentiment categories were defined for further analysis. Having three categories for this study simplifies its interpretation, since the goal is to understand a user’s intention to support, oppose, or be neutral about a candidate without making a difference between extreme and non-extreme sentiments.

Finally, a limitation of MeaningCloud™ is that it can only perform sentiment analysis of up to 10,000 tweets at a time. To address this issue, the datasets for each candidate were stored in Microsoft Excel spreadsheets and sorted by the time of creation from older to newer. Then, the datasets were manually split into subsets of 10,000 tweets. Sentiment analysis was performed in weekly intervals, meaning that chunks of 10,000 tweets were analysed until the end of every week. Figure 5.6 shows the interface of MeaningCloud™ when the Microsoft Excel add-in is used to perform the sentiment analysis.

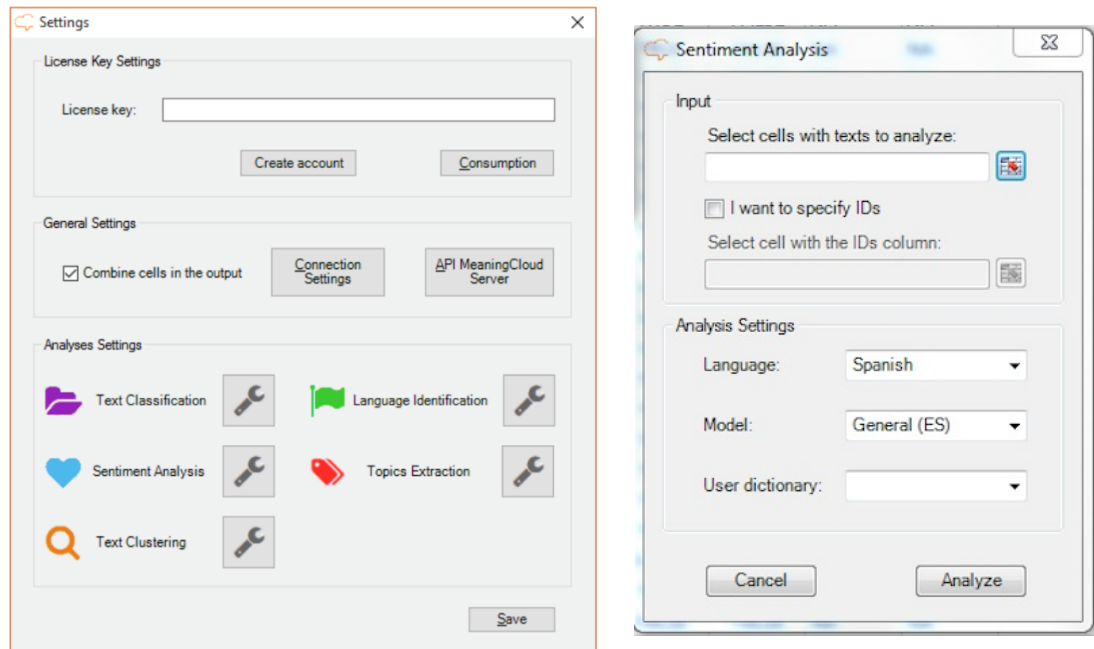


Figure 5.6. Configuration of MeaningCloud™ for sentiment analysis purposes. Input allows the analysis of 10,000 tweets each time taking approximately 45 to 55 minutes to process the data.

In summary, the sentiment analysis conducted in this study involved several choices, some of which can influence the results. First, the choice of keywords for data sampling is perhaps the most crucial issue. Keywords affect directly the search results, and thereby the inclusion/exclusion of tweets for analysis. An important criterion for selecting keywords was to reduce noisiness, as discussed above. For example, by including derogative or offensive names used often to refer to the candidates during the campaign, the performance of the sentiment analysis could be affected. Second, the inclusion of hashtags, emoticons, and URLs in the detection of sentiment is an important aspect of the novelty of this study, and understanding their values to the conveying of sentiment is an issue little addressed in previous research. Instead, previous studies have often discarded these features, arguing that they add no value. And third, the pre-processing approach adopted for the sentiment analysis to transform the data into a machine-readable format to conduct sentiment analysis has the potential to affect results.

Finally, to test if the proposed pre-processing stage has any impact on the performance of the sentiment detection, the sentiment analysis using MeaningCloud™ was also applied using raw tweets, which means tweets in their original format without pre-processing them. Figure 5.7 presents the steps taken from the sentiment analysis.

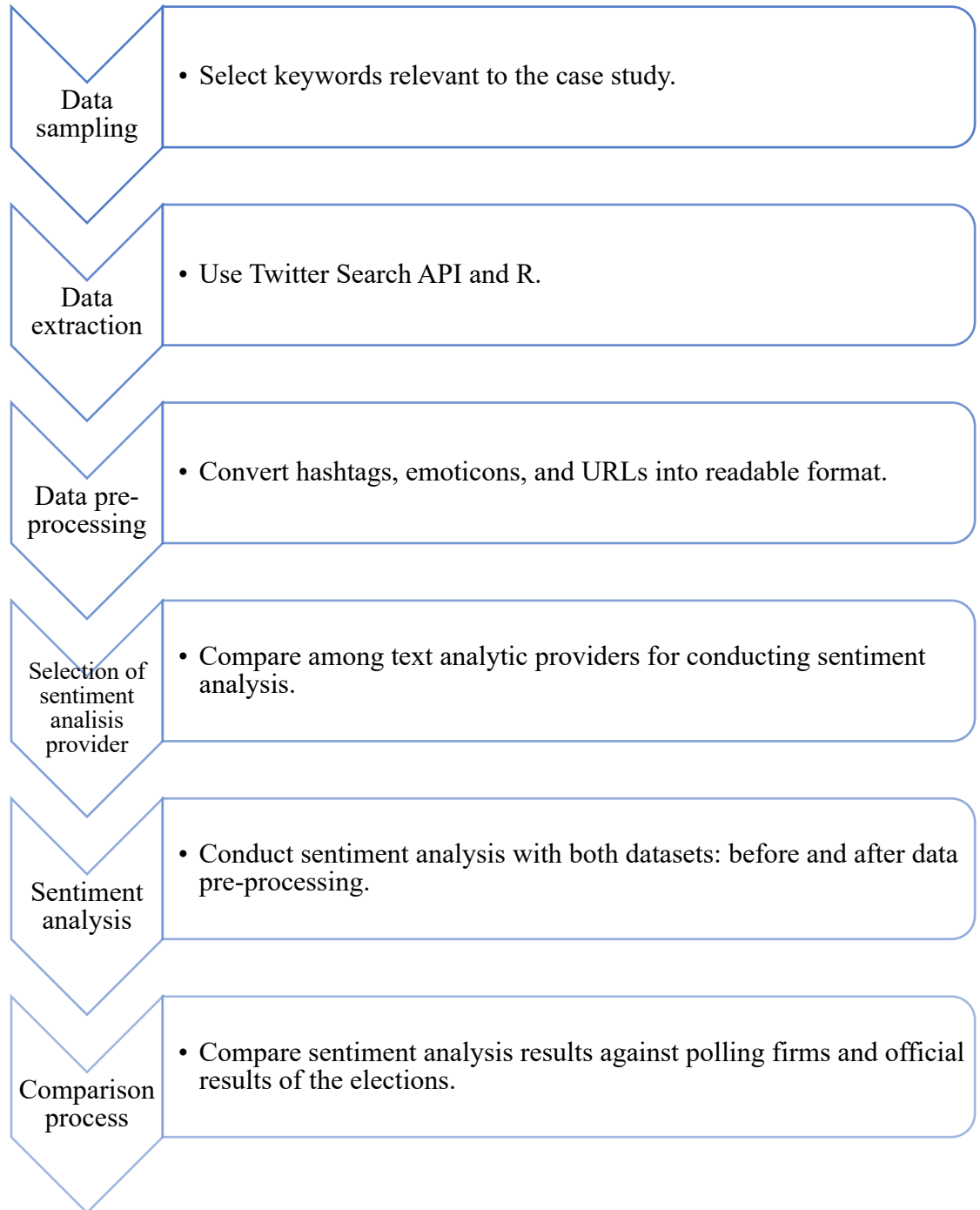


Figure 5.7. Steps taken for the sentiment analysis process.

5.4.6. Identification of influential users

This subsection presents the approach used to identify influential users. Section 5.1 has introduced influential user on Twitter as a user with the ability to engage a large number of other users in a conversation, and thereby to disseminate content quickly and consistently (Chandawarkar, Gould, & Grant Stevens, 2018;

Shattell & Darmoc, 2017). Consistent with this definition and with the approach used in this study, which splits the sentiment analysis in weekly intervals, as stated in Subsection 5.4.5, the definition used for identifying influential users involves two elements: the prevalent mechanism to disseminate content (retweets), and an interval (weekly). These elements emerge from the data as will be explained in the next paragraphs. For this study, *influential users* are those users who produced tweets that were retweeted by many every week. The *most influential users* are the top five users who posted the tweets that obtained the highest number of retweets every week. And, the content of the tweets from these most influential users are what is referred to as the *most retweeted content*. Lastly, the sentiment that the most influential users convey towards the candidates is of particular interest in this study. Further details will be explained in Subsection 5.5.2.

About the issue of interval when investigating influential users, Park, Na, and Moon (2017) have observed that in electoral contexts, the length of influence of a tweet varies depending on its political meaning (e.g. changes in the constitution or economic policy reforms could have longer intervals of influence). However, the literature has said little on how to deal with different lengths of influence, or on an appropriate interval to research influential users on Twitter. A weekly interval was selected in this study because if the interval is too short, e.g. shorter than three days, the analysis could lead to biased results (Ortigosa, Martín, & Carro, 2014). For example, patterns of activity on Twitter might differ between weekdays and weekends (Ilina, 2012). On the other hand, if the interval is too long, some data characteristics may be lost, which could prevent from identifying sporadic influential users (Osmond, 2017). Yet, since other intervals were not tested in this study, further research should be conducted to examine how other intervals can affect the identification of influential users.

Using the same datasets as in the sentiment analysis, the identification of influential users was conducted in two stages. The first one involved identifying the most frequent type of tweet in the datasets; and the second, the identification of users producing the most retweeted content and the sentiment they convey towards the candidates under analysis. Knowing the influential users can help campaigners identify sources of support and opposition, as well as the communities formed around the topics that these users post (Mahmoudi, Yaakub, & Bakar, 2018). Thus, this

approach can provide campaigners with additional knowledge to develop a better targeted communication.

Once the tweets are exported from Twitter to spreadsheet documents, the first step to address is the identification of the most frequent type of tweet, which can fall into one of three categories: “retweet”, “reply”, or “personal” tweet. “Retweet” is a Twitter feature that allows sharing a tweet, and the aim is to reproduce and spread tweets from other Twitter users. These types of tweets can be identified since the tweet is preceded by either the words RT or Retweeted and the Boolean value assigned in the column *isRetweet* is TRUE. “Reply” represents tweets that have been generated as a response to another tweet, and they can be identified in the spreadsheet if the value in the column *replyToSN* is different from NA. Finally, a tweet is considered as “personal” if it does not fall under the two categories mentioned before, this means *isRetweet* is FALSE, and *replyToSN* equals to NA. The second, third, and fourth column in Table 5.4 shows identification of retweet, reply, and personal tweet respectively. The *plyr* package (Wickham, 2011) in R assisted the classification of tweets into these three categories. As will be shown in Subsection 5.5.2, the most frequent type of tweet in these datasets was retweet. Thus, for the purpose of this study, the metric used to define an influential user is retweet.

Table 5.4

Identification of tweets based on the source of the tweet. For illustration purposes, texts of tweets were translated from Spanish to English

Attributes	1	2	3
text	RT @JuventudPais35: Great caravan of our candidates @MashiRafael @Lenin @JorgeGlas @marcelaguinaga @AlainVelez @MichelDoumet @mayracornejo9	@Lenin total triumph Many Blessings.	The political project of #AlianzaPais #Correa, #Glas and #LeninMoreno is to be #Cuba and #Venezuela for the #APes is a paradise https://t.co/Ypcp5cDakK
favorited	FALSE	FALSE	FALSE
favoriteCount	0	0	1
replyToSN	NA	Lenin	NA
created	10/12/2016 00:53:00	09/12/2016 23:28:00	10/12/2016 04:19:00
truncated	FALSE	FALSE	TRUE
replyToSID	NA	NA	NA
id	8.07E+17	8.07E+17	8.07E+17
replyToUID	NA	913131817	NA
statusSource	Twitter for Android	Twitter for Android	Twitter Web Client
screenName	CaroVargas184	LoorEddie	shababaty
retweetCount	21	0	4
isRetweet	TRUE	FALSE	FALSE
retweeted	FALSE	FALSE	FALSE
longitude	NA	NA	NA
latitude	NA	NA	NA

Then, the five most influential users for each of the candidates were identified every week, and their sentiments were manually determined by analysing their previous tweets. These five most influential users accounted for 40.42% of retweets relevant to Lenin Moreno, and 47.07% for Guillermo Lasso. For further details on the identification mechanism of influential users, see Subsection 5.5.2. This stage can provide campaigners with a tool to, for instance, either endorse or harm the credibility of influential users via the candidates' personal accounts or by endorsing others to react to them.

5.5. Results and discussion

The first part of this section covers the sentiment analysis, while the second one focuses on the identification of influential users.

5.5.1. Sentiment analysis results

To determine if the pre-processing stage proposed in this study could enhance the sentiment analysis results, MeaningCloud™ was applied before (using the raw tweets) and after the data pre-processing. After MeaningCloud™ completed the process of sentiment analysis during the sixteen weeks, a simple arithmetic counting procedure was employed where each positive tweet about a candidate was assumed to be one vote. In addition, the numbers of unique Twitter users that produced positive tweets about each candidate were considered for comparison as shown in Table 5.5.

Table 5.5

Numbers of positive tweets and positive Twitter users about the 2017 Ecuadorian Presidential election before and after the pre-processing stage

Sentiment analysis results			Candidates		Total
			L. Moreno	G. Lasso	
Raw tweets	Positive tweets	1 st round	276,529	109,811	386,340
		2 nd round	165,916	110,155	276,071
	Positive users	1 st round	40,378	20,692	61,070
		2 nd round	33,617	30,931	64,548
Pre-processed tweets	Positive tweets	1 st round	286,597	116,040	402,637
		2 nd round	171,871	118,552	290,423
	Positive users	1 st round	41,554	21,453	63,007
		2 nd round	34,346	32,467	66,813

After implementing the pre-processing stage, a number of tweets that had been previously labelled as neutral and negative when analysing the raw datasets, turned into positive. Then, the volume of positive tweets and unique users was higher when conducting sentiment analysis with the pre-processing than with the raw tweets. From the pre-processed tweets, Table 5.5 reveals that in the first round, 63,007 unique Twitter users produced 402,637 tweets that were positive to either Lenin Moreno or Guillermo Lasso. In the second round, 66,813 unique users produced 290,423 tweets favouring either of the two candidates analysed. The disparity in the number of tweets and users in the first and second rounds is firstly due to the difference in time length for each of the rounds (12 and 4 weeks respectively). Secondly, the number of tweets and users that supported candidates other than Lenin Moreno and Guillermo Lasso in the first round (there were 8 candidates in the first round) were (unequally) distributed between the two most voted candidates in the second round.

Table 5.6

Twitter and official results from the 2017 Ecuadorian Presidential election

Results			Candidates		Abs Error*
			L. Moreno	G. Lasso	
Raw tweets	Positive tweets	1 st round	71.58%	28.42%	13.23%
		2 nd round	60.10%	39.90%	8.94%
	Positive Twitter users	1 st round	66.12%	33.88%	7.77%
		2 nd round	52.08%	47.92%	0.92%
Pre- processed tweets	Positive tweets	1 st round	71.18%	28.82%	12.83%
		2 nd round	59.18%	40.82%	8.02%
	Positive Twitter users	1 st round	65.95%	34.05%	7.60%
		2 nd round	51.41%	48.59%	0.25%
Official results		1 st round ¹⁰	58.35%	41.65%	
		2 nd round	51.16%	48.84%	

Source: Official results taken from

<https://resultados2017.cne.gob.ec/frmResultados.aspx>

* The absolute error is the difference between the official results of the 2017 Ecuadorian Presidential election and the results obtained from the sentiment analysis.

¹⁰ Actual official results for the first round were: Lenin Moreno 39.36% and Guillermo Lasso 28.09%. To allow comparison with the numbers of tweets and users, the numbers shown in official results for first round are adjusted after assuming that the two candidates account for 100% of the votes.

Results from the sentiment analysis, before and after the pre-processing stage, and official results from the elections are presented in Table 5.6. The sentiment analysis performed better with the pre-processed tweets than with the raw tweets. This can be seen in the reduced level of error when comparing the election official results against positive tweets and positive unique Twitter users before and after pre-processing for both candidates. From Table 5.6, take for instance the positive Twitter users during the second round. Against the official results, the difference when using raw tweets is 0.92%, while with the pre-processed tweets, this difference shrank to 0.25%. This suggests that the sentiment detection task was slightly improved after pre-processing tweets. Even though the difference between the errors using raw and pre-processed tweets might seem small, the results showed that the pre-processing stage enhanced the ability of MeaningCloud™ to detect the sentiment for the 2017 Ecuadorian Presidential election.

In addition, the time for performing sentiment analysis using the proposed pre-processing method was reduced. Without pre-processing, sentiment analysis for every 10,000 tweets took 55 minutes on average, whereas after pre-processing the average time was 45 minutes. This meant a saving of approximately 1,300 minutes of work. Since raw tweets contain more information to be analysed, noisiness and dimensionality in data are higher than if pre-processed, and thereby the sentiment prediction takes longer. Furthermore, when conducting the sentiment analysis without pre-processing data, MeaningCloud™ experienced system crashes 15 times because the raw tweets included symbols and special characters that the software could not recognise by assuming that these characters were part of a language that was not implemented in the system. Every system-crash implied that the analysis of 10,000 tweets batch suddenly stopped and needed to be restarted. This meant about 825 additional minutes of work (15 times 55 minutes of analysis), which added to the 1,300 minutes mentioned above totals 2,125 minutes. Comparatively, crashes did not happen when using pre-processed data. In terms of runtime of the codes, however, the pre-processing stage took approximately 180 minutes. Therefore, besides improving the sentiment detection, the pre-processing stage reduced the working time and avoided the presence of system crash.

Moreover, the pre-processing stage allowed for a slightly better estimation of the vote share when compared to the raw tweets. Taking the number of users producing tweets with positive sentiment as a proxy of vote choice provided a close

approximation to the official election's results. That is, as shown in Table 5.6 above, the official result was 51.15% and 48.84% for Lenin Moreno and Guillermo Lasso respectively, while the sentiment analysis with the pre-processing task estimated 51.41% and 48.59%. The results with the raw tweets were 52.08% and 47.92%. This apparent gain in accuracy might have its origin in the consideration and approach to coding of hashtags, emoticons, and URLs in the detection of sentiment. Therefore, the small difference between the outcome of the sentiment analysis and the actual election results, especially when using the pre-processing approach introduced in this study, suggests that the proportion of Twitter user accounts producing positive content about a candidate is, in the context of this study, a valid proxy of vote share. This is consistent with the research by Jaidka, Ahmed, Skoric, and Hilbert (2018) and Sang and Bos (2012).

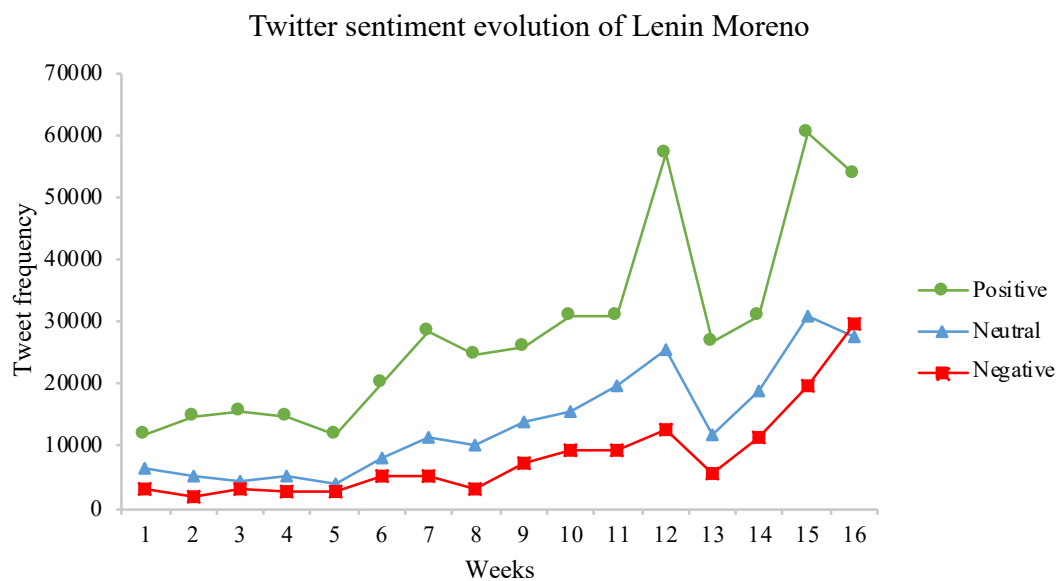


Figure 5.8. Evolution of sentiment on Twitter of the candidate Lenin Moreno during the two phases of elections.

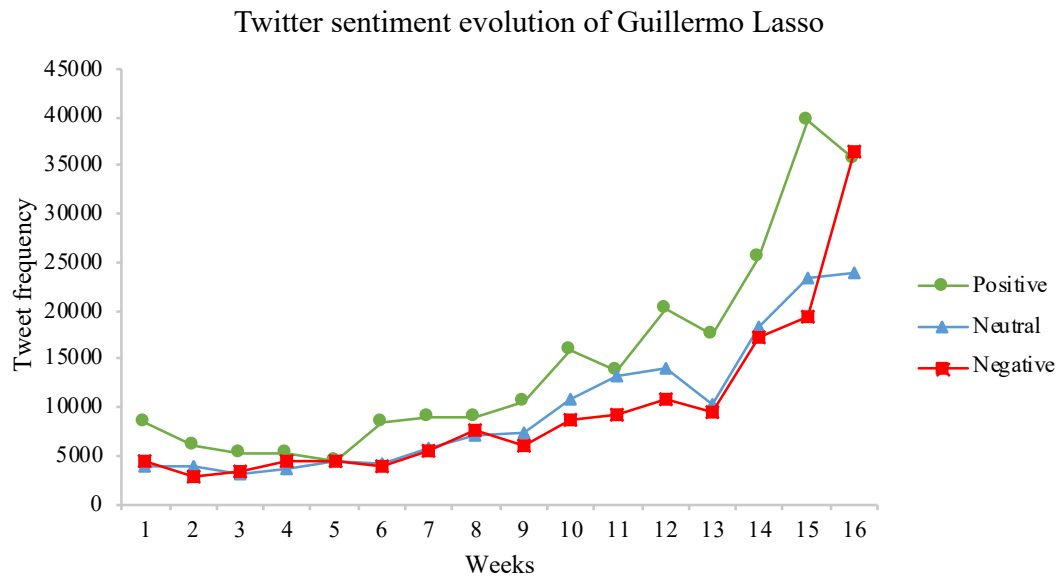


Figure 5.9. Evolution of sentiment on Twitter of the candidate Guillermo Lasso during the two phases of elections.

The weekly evolution of the sentiment of tweets throughout the two phases (16 weeks) is presented in Figure 5.8 and Figure 5.9 for each candidate. The different scales of tweet-frequencies for the two candidates, in Figure 5.8 and Figure 5.9, reflect that users were more prone to tweet positive and neutral tweets about Lenin Moreno. The number of positive tweets reached a peak of 60,565 for Lenin Moreno and 39,624 for Guillermo Lasso, both in week 15. Conversely, users were more prone to tweet content with negative sentiment towards Guillermo Lasso. The figures also show, for each candidate, a pattern among the frequency of tweets by sentiment, whereby when the positive tweets increase or decrease from one week to the next, so do, in most cases, the negative and neutral. Likewise, the figures show that, as the campaign went forward to the voting days in weeks 12 and 16, the overall number of tweets increased.

Figure 5.8 shows a sharp decline in the positive tweets during week 12 and a slight rise in negative tweets in week 13 relevant to Lenin Moreno, which returned to the growing trend in the following two weeks. In the case of Guillermo Lasso as shown in Figure 5.9, the tendency of tweet volume was foremost increasing throughout the campaign, and more abruptly during the second round. This can be associated with the fact that in the second round he was the only opponent candidate. Nevertheless, a sharp drop in positive tweets against a severe increase in negative tweets in the last week prior to the election date may explain why polling firms failed

to anticipate the results, as seen below in Table 5.7. This drop of positive tweets during the last week of the campaign can be associated to a video showing Guillermo Lasso supporters supposedly attacking an elderly person¹¹. This event, which was widely retweeted among users opposing Guillermo Lasso, triggered the increase in negative sentiment towards this candidate.

Table 5.7

Pre-vote polls and official results from the 2017 Ecuadorian Presidential election

Candidates	Market		Cedatos		Opinión Pública		Perfiles de Opinión	
	1st round	2nd round	1st round	2nd round	1st round	2nd round	1st round	2nd round
L. Moreno	28.49%	52.08%	32.30%	52.41%	34.20%	57.48%	35.00%	57.59%
G. Lasso	18.29%	47.92%	21.50%	47.59%	18.20%	42.52%	16.00%	42.41%
Error:								
L. Moreno	-10.87%	0.92%	-7.06%	1.25%	-5.16%	6.32%	-4.36%	6.43%
G. Lasso	-9.80%	-0.92%	-6.59%	-1.25%	-9.89%	-6.32%	-12.09%	-6.43%

Source: Summary of polls in Newspaper “El Universo”¹². These results correspond to the last polls conducted before the elections took place in the two rounds.

The firms *Market*, *Cedatos*, *Opinión Pública*, and *Perfiles de Opinión* were the polling firms authorised by the CNE to conduct polls and report their results in the media. To obtain these results, these polling firms conducted pre-vote polls mainly by interviewing voters, in samples that ranged between 2,000 and 3,000 during a timeline of 2 to 3 days during the mid of March 2018¹². The mean error of the polling firms, in terms of the official results, was 5.98%. For the first round, the mean error was 8.23%, while for the second round this error accounted for 3.73% (see Table 5.7). In comparison with the Twitter users’ sentiment analysis results (first round 7.6% and second round 0.25%), this study shows that the errors of the results of the sentiment analysis after the pre-processing stage with respect to the official results were smaller than those of the polling firms. And, even the results from the sentiment analysis based on raw tweets were closer to the official ones than the official polling firms.

5.5.2. Identification of influential users

From the approximately 1.3 million tweets extracted during the elections, it can be observed in Table 5.8 that 84.38% of tweets fell under the retweet category,

¹¹ <https://twitter.com/PABLOJION/status/848258468045885441>

¹² <http://www.eluniverso.com/noticias/2017/03/22/nota/6101566/encuestadoras-dan-ultimo-reporte-intencion-voto>

8.53% were composed of replies to other Twitter user accounts, and 6.64% were personal tweets.

Table 5.8

Proportion of tweets generated by different Twitter users regarding the two candidates

Round	Candidates	Total tweets	Number of tweets		
			Retweet	Reply	Personal
1 st	Lenin Moreno	482,318	431,177	26,063	25,078
	Guillermo Lasso	269,252	216,504	35,996	16,752
2 nd	Lenin Moreno	328,259	284,483	18,741	25,035
	Guillermo Lasso	276,899	229,431	26,653	20,815
	Total	1,356,728	1,161,595	107,453	87,680

Since the clear majority of tweets were retweets, the next step was the identification of the Twitter user accounts which had the highest number of retweets every week to be considered as the most influential users. The number of retweets can provide a glimpse about the popularity of both content and users.

In the dataset containing information about Lenin Moreno, it can be observed that his official Twitter communicational and personal accounts, *@VamosLenin* and *@Lenin*, were consistently the most influential users throughout the 16 weeks. In the case of the dataset of Guillermo Lasso, his own personal account, *@LassoGuillermo*, was the most influential one in the whole campaign. Further, the 5 most influential users during the 16 weeks generated almost 44% of the total number of retweets. From the 80 most influential users, it can be noted that 12 unique Twitter user accounts generated the most retweeted content concerning Lenin Moreno, while for Guillermo Lasso was generated by 28 different Twitter user accounts. Additionally, in the dataset of Lenin Moreno, 96% of the retweets came from tweets posted by either Lenin Moreno himself or by any of the ten other most influential users that supported him throughout the campaign. Thus, the majority of retweets in this dataset had a positive sentiment towards Lenin Moreno. Only the remaining 4% of retweets conveyed negative sentiments, which came from tweets generated by Guillermo Lasso. In the dataset about Guillermo Lasso, on the other hand, 39% of the retweets came from his own tweets or from any of the nine other users supporting him consistently. 59% of the retweets in this dataset had a negative sentiment about Guillermo Lasso and were generated by sixteen users. The remaining 2% had

conveyed neither positive nor negative sentiments towards him. Details of the most influential users per week relevant to both candidates are presented in Table 5.9 and Table 5.10.

As mentioned in Subsection 5.4.6, the opportune identification of influential users and their sentiments towards the candidates can help to identify the communities formed around influential users and thereby develop a better targeted communication. For example, if the influential users are supportive towards the candidates, politicians can promote and retweet their content among followers, or engage them more deeply in the campaign to continue spreading viral positive tweets. When influential users have instead been unsupportive, the literature has identified different approaches in which it has been addressed. In some cases, the communication strategy has focused on harming the credibility of the negative influential users among his/her followers by using criticism or mocking tactics (Lee & Lim, 2016). Others have bought supportive followers to reduce the perception of negativity (Confessore, Dance, Harris, & Hansen, 2018). And others have blocked, muted, or reported an account to have it suspended. However, sometimes doing nothing has been a valid strategy to avoid giving the influencer user further publicity.

Table 5.9

*Most influential users from those tweeting about Lenin Moreno**

Week	Username	Lenin	leninmorenospais	VamosLenin	35pais	micheldoumet
1	Retweets	3,422	2,195	1,855	1,200	411
Week	Username	Lenin	leninmorenospais	VamosLenin	maximaccion35ap	35pais
2	Retweets	5,161	1,194	1,180	594	543
Week	Username	Lenin	VamosLenin	leninmorenospais	35pais	marialevicuna
3	Retweets	3,278	3,144	1,632	1,013	457
Week	Username	VamosLenin	Lenin	leninmorenospais	35pais	LassoGuillermo
4	Retweets	5,049	4,534	1,631	469	398
Week	Username	VamosLenin	Lenin	leninmorenospais	35pais	humanistasec
5	Retweets	4,105	3,465	1,069	562	473
Week	Username	VamosLenin	Lenin	leninmorenospais	35pais	LassoGuillermo
6	Retweets	6,704	6,491	1,772	817	758
Week	Username	VamosLenin	Lenin	leninmorenospais	jorgeglas	35pais
7	Retweets	10,610	5,941	2,529	1,491	1,209
Week	Username	VamosLenin	Lenin	jorgeglas	leninmorenospais	notilenin
8	Retweets	6,772	5,105	2,444	1,816	857
Week	Username	VamosLenin	Lenin	humanistasec	leninmorenospais	jorgeglas
9	Retweets	7,401	5,167	2,540	1,422	1,190
Week	Username	VamosLenin	Lenin	leninmorenospais	humanistasec	35pais
10	Retweets	10,029	5,258	2,848	2,359	1,424
Week	Username	VamosLenin	Lenin	micheldoumet	jorgeglas	leninmorenospais
11	Retweets	8,576	4,800	3,135	2,557	2,368
Week	Username	VamosLenin	Lenin	leninmorenospais	micheldoumet	jorgeglas
12	Retweets	13,292	11,519	3,854	3,180	2,642
Week	Username	VamosLenin	Lenin	35pais	leninmorenospais	35apd3gye
13	Retweets	6,127	3,842	2,350	1,626	1,494
Week	Username	VamosLenin	Lenin	35pais	micheldoumet	leninmorenospais
14	Retweets	8,055	4,006	2,821	2,504	1,547
Week	Username	VamosLenin	Lenin	micheldoumet	leninmorenospais	35pais
15	Retweets	9,105	6,679	3,926	3,728	3,648
Week	Username	Lenin	VamosLenin	LassoGuillermo	35pais	leninmorenospais
16	Retweets	10,277	6,351	6,307	2,749	2,193

* Orange highlight colour refers to the Twitter user accounts that are negative towards the candidate under analysis. If not highlighted, a Twitter user account is assumed to be supportive of the candidate under analysis.


Table 5.10

Most influential users from those tweeting about Guillermo Lasso^o

Week	Username	LassoGuillermo	lolacienfuegos	lospoliticosec	cpi_consultores	notcreolasso
1	Retweets	7,390	257	245	205	200
Week 2	Username	LassoGuillermo	parischiquitoo	lolacienfuegos	avecillalibre	notcreolasso
	Retweets	6,280	221	148	145	124
Week 3	Username	LassoGuillermo	somosmasec	lolacienfuegos	politiquerosec	parischiquitoo
	Retweets	4,659	341	284	281	265
Week 4	Username	LassoGuillermo	lospoliticosec	lolacienfuegos	somosmasec	leninmorenospais
	Retweets	4,824	603	352	264	260
Week 5	Username	LassoGuillermo	parischiquitoo	lospoliticosec	lolacienfuegos	notcreolasso
	Retweets	3,522	503	460	414	231
Week 6	Username	LassoGuillermo	lolacienfuegos	leninmorenospais	parischiquitoo	notcreolasso
	Retweets	6,831	560	415	225	222
Week 7	Username	LassoGuillermo	notcreolasso	lolacienfuegos	parischiquitoo	vivianassange
	Retweets	5,792	879	341	288	283
Week 8	Username	LassoGuillermo	lospoliticosec	notcreolasso	daloel10	VamosLenin
	Retweets	5,384	2,103	1,024	421	357
Week 9	Username	LassoGuillermo	notcreolasso	lospoliticosec	anonymous_ec	tioelmoi
	Retweets	5,365	1,105	423	367	337
Week 10	Username	LassoGuillermo	notcreolasso	lospoliticosec	sipodemosec	soldadaturca
	Retweets	11,951	1,027	906	758	646
Week 11	Username	LassoGuillermo	notcreolasso	mashirafael	eluniversocom	lospoliticosec
	Retweets	9,205	2,270	1,294	474	459
Week 12	Username	LassoGuillermo	fevillavicencio	wikileaks	notcreolasso	andrespaezsec
	Retweets	15,871	2,054	708	603	526
Week 13	Username	LassoGuillermo	eluniversocom	soldadaturca	lahistoriaec	notcreolasso
	Retweets	6,717	1,236	1,108	1,033	714
Week 14	Username	LassoGuillermo	el_telegrafo	eluniversocom	bloglibrecuador	padrejoselamar
	Retweets	12,224	3,467	1,814	1,350	1,299
Week 15	Username	LassoGuillermo	el_telegrafo	eluniversocom	mashirafael	pljv7
	Retweets	24,601	3,226	2,192	1,255	1,085
Week 16	Username	LassoGuillermo	el_telegrafo	cnnee	andrespaezsec	lahistoriaec
	Retweets	26,325	1,635	1,614	1,607	1,459

Analysing the retweeting behaviour, it can also be identified that the most influential users were mainly composed of verified politicians, unverified Twitter user accounts supporting or opposing each candidate, and verified parties' accounts. A verified account is one that Twitter has certified as authentic (Caruccio, Desiato, &

^o Orange highlight colour refers to the Twitter user accounts that are negative towards the candidate under analysis, while yellow is for those neutral. If not highlighted, a Twitter user account is assumed to be supportive of the candidate under analysis.

Polese, 2018) and has a blue badge  next to the account's profile¹³. On the other hand, accounts without the blue badge are considered not verified, which are called unverified accounts. Surprisingly, for these elections, traditional media only constituted a small portion of the most influential users, reflecting that the hegemony which the traditional media have had in the past was somehow disintermediated by other Twitter user accounts in political matters.

5.5.3. *Evolution of followers during the campaign*

This section is relevant because it shows that followership of candidates is not a consistent proxy for support. Influence and popularity of Twitter users have been commonly measured in terms of number of followers (Cha et al., 2010; Kwak et al., 2010). Based on this, the numbers of followers and followees of the two candidates were collected at the end of every week during the two rounds. Figure 5.10 shows the evolution of followers. It can be observed that Guillermo Lasso had more followers during the whole campaign than Lenin Moreno. However, comparing the evolution in the number of followers for the two candidates, Guillermo Lasso added 63,000 new followers until the last day of the campaign, whereas Lenin Moreno added 219,500 new followers during the 16 weeks (3:5 proportion in comparison with Guillermo Lasso). Therefore, the number of followers of Lenin Moreno noticeably rose during the weeks. If merely the number of followers was a measure of influence, Guillermo Lasso would have outperformed Lenin Moreno as influential. However, influence in terms of the number of followers during electoral campaigns needs to be carefully addressed since there might be the intervention of fake or bot followers, as specific complaints emerged about this election (Confessore et al., 2018). In addition, the results generated from a correlation coefficient test between the number of retweets and followers showed that for Lenin Moreno, no correlation was found between these two variables ($r = 0.0360$, $p = 0.8946$), while for Guillermo Lasso there was a positive one ($r = 0.7182$, $p = 0.0017$), as shown in Figure 5.11.

Concerning the number of Twitter user accounts the two candidates followed during the campaign, there was no change in the number of followees for either candidate. Lenin Moreno started the first week following 25 accounts and ended with 26 followees. Likewise, Guillermo Lasso followed 1,453 during the first week, and by the end of campaign he followed 1,457 accounts.

¹³ <https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts>

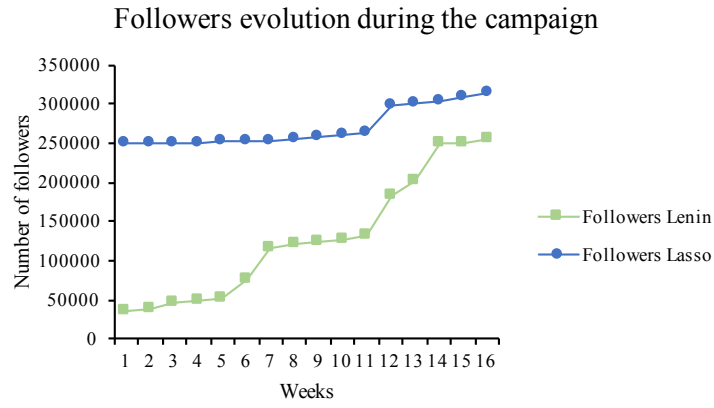


Figure 5.10. Evolution of followers of the two candidates during the presidential campaign.

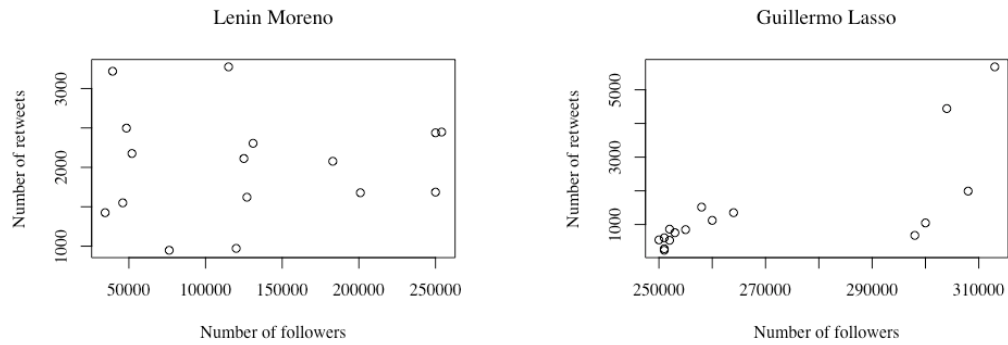


Figure 5.11. Correlation plots between number of retweets and followers for both candidates.

5.6. Conclusion

This study has presented two novel approaches for data analysis using Twitter as its data source. The first one is related to data pre-processing for sentiment analysis purposes, where a new approach for conducting data pre-processing is introduced. The approach integrates the analysis of features of tweets, namely hashtags, emoticons, and URLs, that have been traditionally discarded since they have been considered of little value for sentiment detection purposes in previous research. Regarding hashtags, they were separated into independent words to make the information they contain readable for a sentiment analysis provider. In the case of emoticons, these were replaced with text because they can abstract the sentiment of a tweet. URLs were manually analysed since they followed a power-law distribution, which allowed to perform a precise manual sentiment analysis of the few most

frequent URLs, rather than relying on the results of the automatic analysis of all URLs.

The second novel approach concerns the identification of influential users, which is presented using a two-stage approach to weekly identify who are the users posting tweets that obtain the highest number of retweets. The aim of the first stage was to identify the most frequent type of tweet found in the datasets. For the second stage, the aim was to identify the most influential users on a weekly basis, which is comprised by those that obtained the highest number of retweets as explained in Subsection 5.4.6. Unlike the previous research, this approach is meant to identify influential users from the evidence of the data instead of in an intuitive way, and to enable campaigners to timely spot sources of support and opposition on Twitter.

The study relied on the Twitter data produced during the official campaigning period towards the 2017 Ecuadorian Presidential election. To validate the value of the approaches here presented, the study used a combination of official results and reports of authorised vote share polling firms.

Three key findings are presented. First, the integration and coding of hashtags, emoticons, and URLs, enhanced the performance of the sentiment analysis tool by showing more accurate results than when raw tweets were used for analysis. This means that sentiment analysis became less prone to error, which for this study refers to the difference between the official results from the election and the sentiment analysis results, as shown in Subsection 5.5.1. Also, pre-processing allowed for gains in the time performance of the sentiment analysis and reduced the occurrence of system crashes. These features make pre-processing an important task to consider when conducting sentiment analysis based on Twitter data.

Second, by conducting sentiment analysis with the pre-processed data, this study has shown that the number of unique Twitter users supporting a candidate during a political campaign can be used as a proxy for vote share, which can be more accurate than traditional polls. Even when it is claimed that sentiment models based on Twitter data are not representative of overall public opinion (Mellon & Prosser, 2017; Mitchell & Hitlin, 2013; Schoen et al., 2013; Sinnenberg et al., 2017), this study reveals that the sentiment of tweets is a valid real-time indicator of voters' preferences, which is also supported by previous studies (Borondo et al., 2012; Caldarelli et al., 2014; Tumasjan et al., 2010).

Third, alongside a few users that remained most influential consistently throughout the campaign, several most influential users also emerged momentarily, which could have affected the sentiment. In this regard, campaigners need to be aware of who these users and their networks might be. To address this issue, this study proposed a two-stage approach to identify influential users in weekly intervals as explained in Subsection 5.4.6. The other relevant choice was defined by influential users in terms of their ability to produce the most retweeted content. The reason for this was that the most frequent type of tweet relevant to the campaign was retweets.

Therefore, the approaches presented in this study can enhance the accuracy of Twitter-based sentiment analysis, deliver an indication of the vote share during an electoral campaign, and identify influential users periodically. This is useful, for example, for monitoring changes in the sentiment towards candidates of an electoral race, detecting emerging supporters and opposers to candidates, identify possible sources of change in sentiment towards candidates, and timely address concerns that users might have about the candidates.

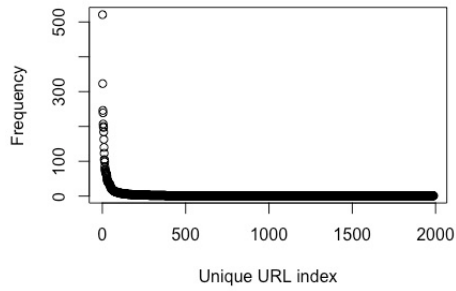
Concerning limitations, there are three in this study worth mentioning. First, the fact that users can either produce large numbers of tweets with one account or have more than one Twitter user account producing tweets can affect the ability of sentiment analysis tools to detect the overall sentiment accurately. Both situations can lead to under- or overestimating sentiment towards a candidate. Second, there is still little development of sentiment analysis tools that work in languages different from English, as was the case in this study. Notwithstanding, MeaningCloudTM performed well in detecting the sentiment of tweets in Spanish. Third, this study did not consider the presence of fake users or bots, nor actual users producing fake news, which can influence users' perceptions. Despite these limitations, the approaches here presented have proven fruitful for the case. This was validated when comparing the ability of the sentiment analysis tool to detect vote share with the official results and those of the polling firms.

Finally, in terms of further research, it is important to raise awareness about the shortage of literature that addresses what it takes for a sample to be significant on Twitter. To what extent do the near 1.3 million tweets relevant to Ecuadorian politics posted by 140,617 users embody a significant sample of the Ecuadorian electorate? Hence, a random sample of likely voters is an unrealistic aim on Twitter, because there is no way to determine if users that produce relevant content are actual voters. It

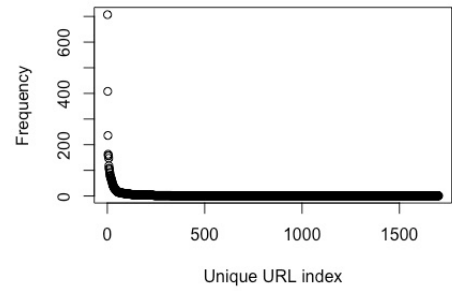
seems, however, that conventional statistical principles of sampling might not be of high relevance here, for which, in the face of the evident ability to detect the sentiment, future research should focus on the principles that enable prediction of electoral outcomes from Twitter content. Finally, the analysis of the scope of the features used for data pre-processing could be individually analysed to determine to what extent they can improve sentiment analysis results.

Appendix A-5.

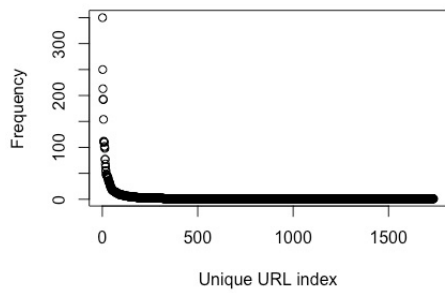
Week 1



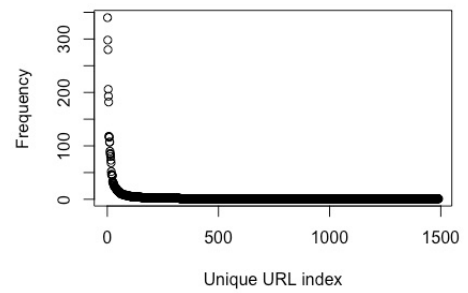
Week 2



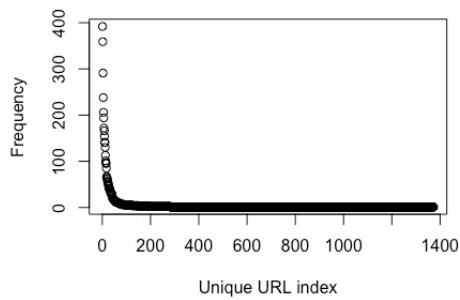
Week 3



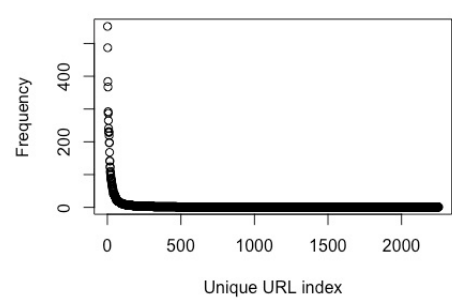
Week 4



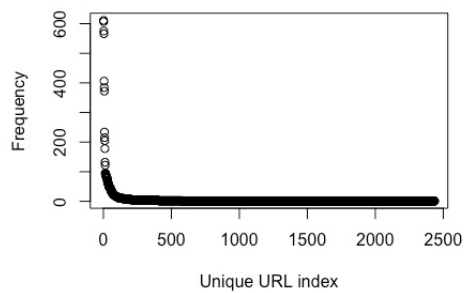
Week 5



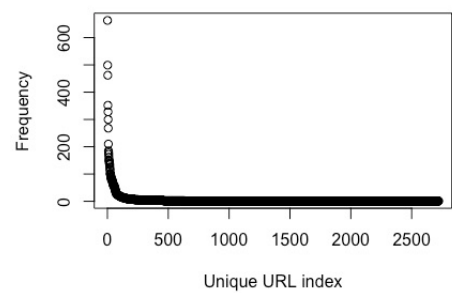
Week 6



Week 7



Week 8



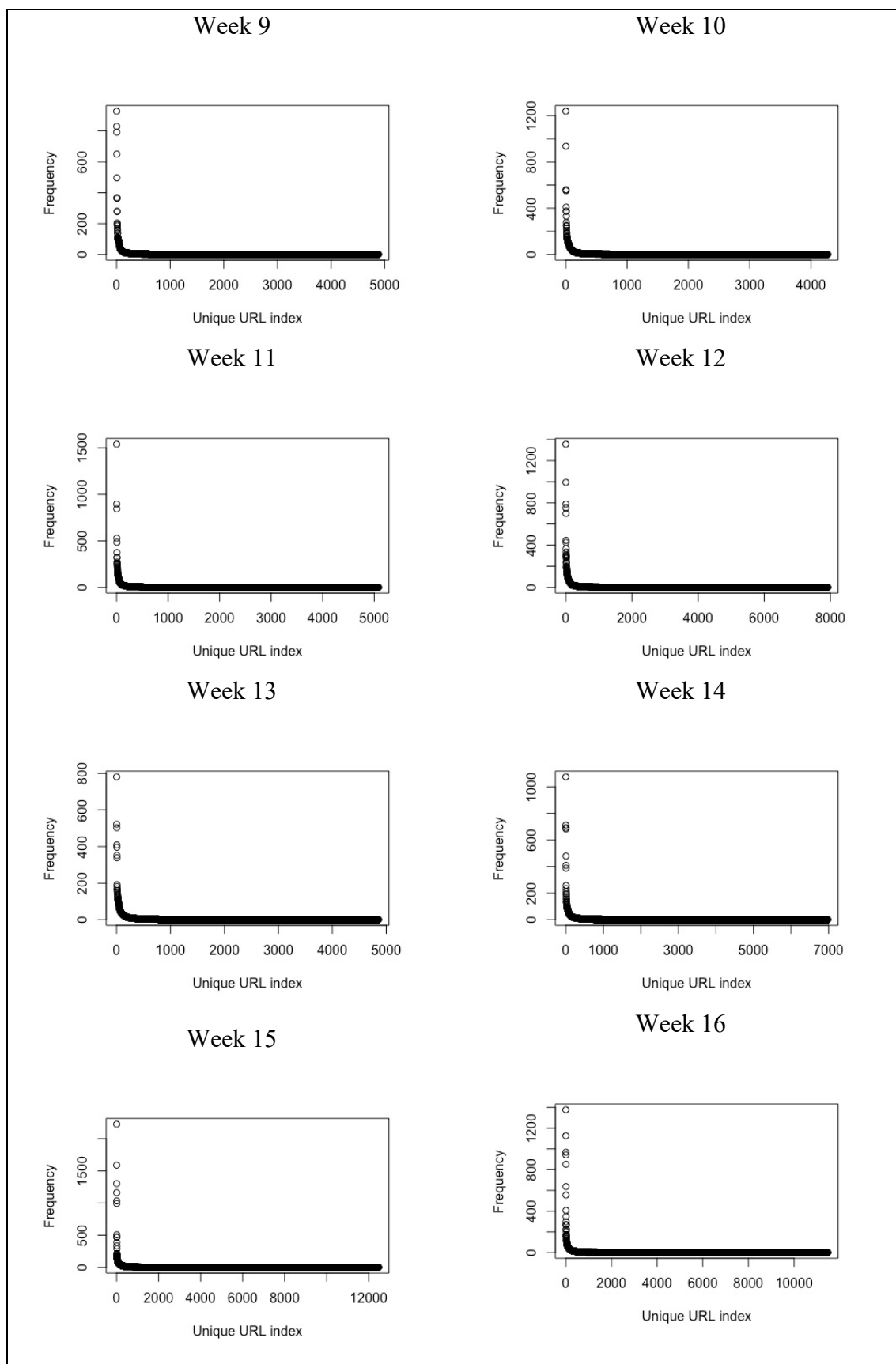
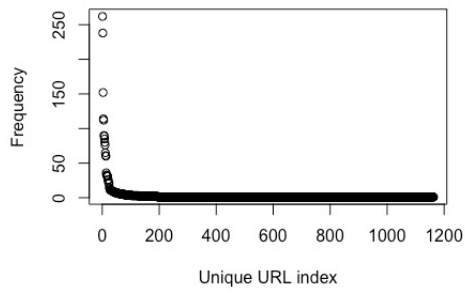
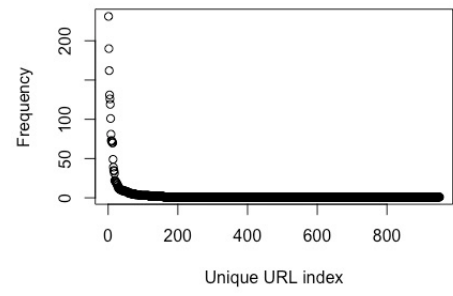


Figure A-5.1. Weekly breakdown of the frequency plot of URLs found in tweets about Lenin Moreno.

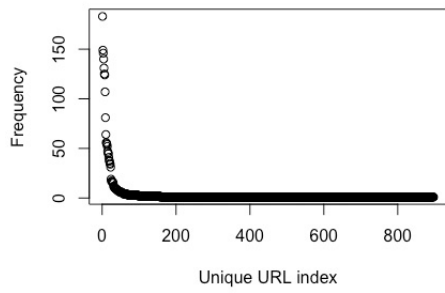
Week 1



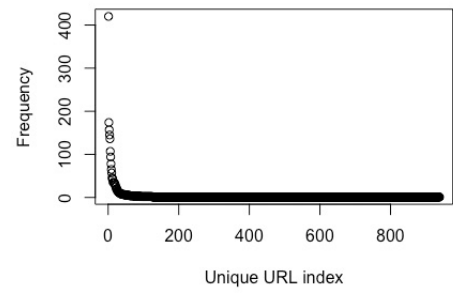
Week 2



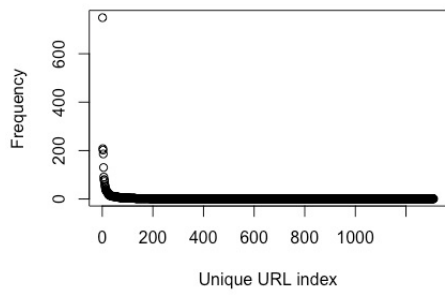
Week 3



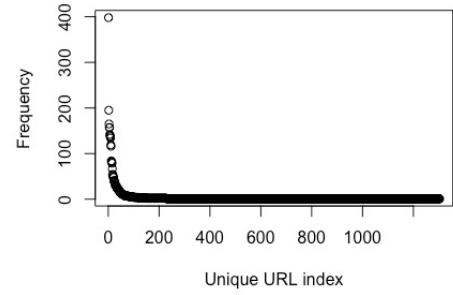
Week 4



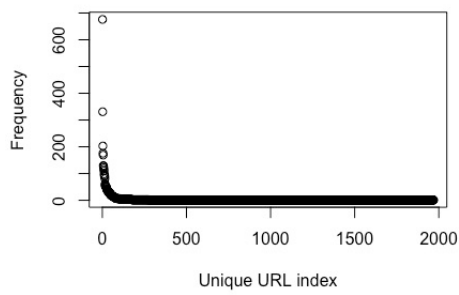
Week 5



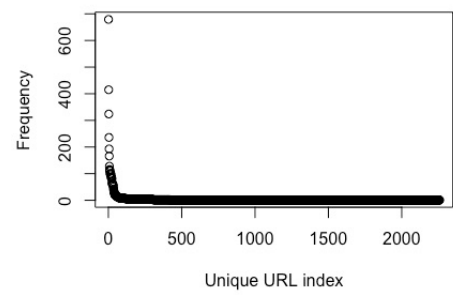
Week 6



Week 7



Week 8



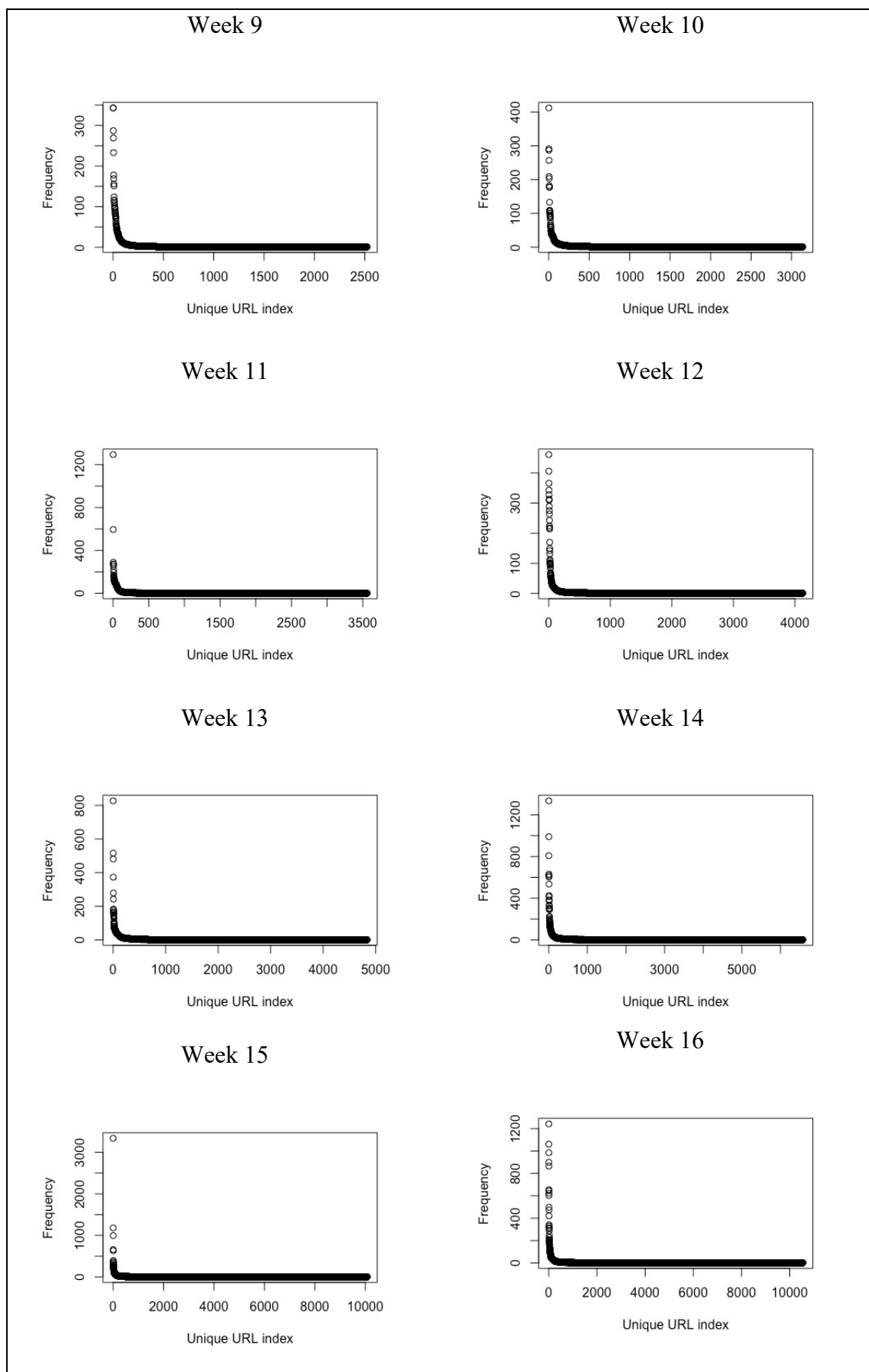


Figure A-5.2. Weekly breakdown of the frequency plot of URLs found in tweets about Guillermo Lasso.

References

- Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Languages in Social Media* (pp. 30-38). Association for Computational Linguistics, Portland, Oregon, US.
- Al Hamoud, A., Alwehaibi, A., Roy, K., & Bikdash, M. (2018). Classifying political tweets using Naïve Bayes and Support Vector Machine. In *31st International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems* (pp. 736-744). Springer, Montreal, Quebec, Canada.
- Alp, Z. Z., & Ögüdücü, Ş. G. (2018). Influence factorization for identifying authorities in Twitter. *Journal of Knowledge-Based Systems*, 163, 944-954.
- Amaghlobeli, N. (2012). Linguistic features of typographic emoticons in SMS discourse. *Journal of Theory and Practice in Language Studies*, 2(2), 348.
- Anderson, B. (2017). Tweeter-in-Chief: A content analysis of president Trump's tweeting habits. *Elon Journal of Undergraduate Research in Communications*, 8(2), 36-47.
- Anjaria, M., & Guddeti, R. M. R. (2014). A novel sentiment analysis of social networks using supervised learning. *Journal of Social Network Analysis and Mining*, 4(181), 1-15.
- Aral, S., & Walker, D. (2012). Identifying influential and susceptible members of social networks. *Journal of Science*, 337(6092), 337-341.
- Aral, S., & Walker, D. (2014). Tie strength, embeddedness, and social influence: A large-scale networked experiment. *Journal of Management Science*, 60(6), 1352-1370.
- Aramaki, E., Maskawa, S., & Morita, M. (2011). Twitter catches the flu: Detecting influenza epidemics using Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1568-1576). Association for Computational Linguistics, Edinburgh, Scotland, UK.
- Arcuri, L., Castelli, L., Galdi, S., Zogmaister, C., & Amadori, A. (2008). Predicting the vote: Implicit attitudes as predictors of the future behaviour of decided and undecided voters. *Journal of Political Psychology*, 29(3), 369-387.
- Asur, S., & Huberman, B. (2010). Predicting the future with social media. In *2010 International Conference on Web Intelligence and Intelligent Agent Technology* (pp. 492-499). IEEE, Los Alamitos, California, US.
- Ausserhofer, J., & Maireder, A. (2013). National politics on Twitter: Structures and topics of a networked public sphere. *Journal of Information, Communication & Society*, 16(3), 291-314.
- Auvinen, A.-M. (2012). Social media: The new power of political influence: Suomen Toivo Think Tank. Centre for European Studies.
- Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 2010 Conference on Language Resource and Evaluation* (pp. 2200-2204). European Languages Resources Association, Valletta, Malta.
- Bae, Y., & Lee, H. (2012). Sentiment analysis of Twitter audiences: Measuring the positive or negative influence of popular twitterers. *Journal of the American Society for Information Science and Technology*, 63(12), 2521-2535.
- Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone is an influencer: Quantifying influence on Twitter. In *Proceedings of the International Conference on Web Search and Data Mining* (pp. 65-74). ACM, Hong Kong.

- Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Journal of Political Analysis*, 23(1), 76-91.
- Barbosa, L., & Feng, J. (2010). Robust sentiment detection on Twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 36-44). Association for Computational Linguistics, Beijing, China.
- Bermingham, A., & Smeaton, A. (2011). On using Twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis: Where AI meets Psychology* (pp. 2-10). Workshop at the International Joint Conference for Natural Language Processing. Association for Computational Linguistics, Chiang Mai, Thailand.
- Bilro, R. G., Loureiro, S. M. C., & Guerreiro, J. (2018). Analysing customer engagement on social network platforms devoted to tourism and hospitality. In *2018 Conference on Global Marketing* (pp. 239-240). Global Alliance of Marketing and Management Associations, Tokyo, Japan.
- Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1-8.
- Borondo, J., Morales, A., Losada, J. C., & Benito, R. M. (2012). Characterizing and modelling an electoral campaign in the context of Twitter: 2011 Spanish Presidential election as a case study. *Journal of Chaos*, 22(2), 023138-023137.
- Bovet, A., Morone, F., & Makse, H. A. (2016). Predicting election trends with Twitter: Hillary Clinton versus Donald Trump. *arXiv preprint arXiv:1610.01587*.
- Boyd, D., Golder, S., & Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In *2010 International Conference on System Sciences* (pp. 1-10). IEEE, Honolulu, Hawaii, US.
- Bruns, A. (2012). How long is a tweet? Mapping dynamic conversation networks on Twitter using Gawk and Gephi. *Journal of Information, Communication & Society*, 15(9), 1323-1351.
- Budiharto, W., & Meiliana, M. (2018). Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis. *Journal of Big Data*, 5(1), 51-61.
- Burgess, J., & Bruns, A. (2012). (Not) The Twitter election: The dynamics of the #ausvotes conversation in relation to the Australian media ecology. *Journalism Practice*, 6(3), 384-405.
- Burnap, P., Gibson, R., Sloan, L., Southern, R., & Williams, M. (2016). 140 characters to victory? Using Twitter to predict the UK 2015 General election. *Journal of Electoral Studies*, 41(2016), 230-233.
- Caldarelli, G., Chessa, A., Pammolli, F., Pompa, G., Puliga, M., Riccaboni, M., & Riotto, G. (2014). A multi-level geographical study of Italian political elections from Twitter data. *Journal of PLOS ONE*, 9(5), e95809.
- Cambria, E. (2013). An introduction to concept-level sentiment analysis. In *Mexican International Conference on Artificial Intelligence* (pp. 478-483). Springer, Mexico City, Mexico.
- Cameron, M. P., Barrett, P., & Stewardson, B. (2016). Can social media predict election results? Evidence from New Zealand. *Journal of Political Marketing*, 15(4), 416-435.
- Caruccio, L., Desiato, D., & Polese, G. (2018). Fake account identification in social networks. In *2018 International Conference on Big Data* (pp. 5078-5085). IEEE, Seattle, US.

- Carvajal, A. (2017). Las redes sociales también son "tarima" de los presidenciables [Social media platforms are also the "pallets" of presidential candidates]. *El Comercio*. Retrieved on March 4th, 2017 from <http://www.elcomercio.com/actualidad/redes-sociales-son-tarima-presidenciables.html>.
- Cawsey, T., & Rowley, J. (2016). Social media brand building strategies in B2B companies. *Journal of Marketing Intelligence & Planning*, 34(6), 754-776.
- Ceron, A., Curini, L., & Iacus, S. M. (2015). Using sentiment analysis to monitor electoral campaigns: Method matters - Evidence from the United States and Italy. *Journal of Social Science Computer Review*, 33(1), 3-20.
- Ceron, A., & d'Adda, G. (2016). E-campaigning on Twitter: The effectiveness of distributive promises and negative campaign in the 2013 Italian election. *Journal of New Media & Society*, 18(9), 1935-1955.
- Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring user influence in Twitter: The million follower fallacy. In *Proceedings of the Fourth International AAAI Conference on Web and Social Media* (pp. 10-18). AAAI Press, Menlo Park, California, US.
- Chandawarkar, A. A., Gould, D. J., & Grant Stevens, W. (2018). The top 100 social media influencers in plastic surgery on Twitter: Who should you be following? *Aesthetic Surgery Journal*, 38(8), 913-917.
- Cheong, Y. P., & Gupta, R. (2005). Experimental design and analysis methods for assessing volumetric uncertainties. *SPE Journal*, 10(03), 324-335.
- Choi, Y., & Lee, H. (2017). Data properties and the performance of sentiment classification for electronic commerce applications. *Journal of Information Systems Frontiers*, 19(5), 993-1015.
- Chy, A. N., Ullah, M. Z., & Aono, M. (2015). Combining temporal and content aware features for microblog retrieval. In *2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications* (pp. 1-6). IEEE, Chonburi, Thailand.
- Clark, T. (2015). New research suggests why General election polls were so inaccurate. *The Guardian*. Retrieved on May 3th, 2016 from <http://www.theguardian.com/politics/2015/nov/13/new-research-general-election-polls-inaccurate>.
- Cody, E. M., Reagan, A. J., Dodds, P. S., & Danforth, C. M. (2016). Public opinion polling with Twitter. *arXiv preprint arXiv:1608.02024*.
- Confessore, N., Dance, G., Harris, R., & Hansen, M. (2018). The follower factory. Retrieved on February 2nd, 2018 from <https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html?smid=tw-nytimes&smtyp=cur>.
- Conover, M., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F., & Flammini, A. (2011a). Political polarization on Twitter. In *Proceedings of the Fifth International AAAI Conference on Web and Social Media* (pp. 89-96). AAAI Press, Barcelona, Spain.
- Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., & Menczer, F. (2011b). Predicting the political alignment of Twitter users. In *Proceedings of the International Conference on Privacy, Security, Risk and Trust, and on Social Computing* (pp. 192-199). IEEE, Boston, Massachusetts, US.
- Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical*

- Reality* (pp. 51-60). Association for Computational Linguistics, Baltimore, Maryland, US.
- Cossu, J.-V., Dugué, N., & Labatut, V. (2015). Detecting real-world influence through Twitter. In *2015 Second European Network Intelligence Conference* (pp. 83-90). IEEE, Karlskrona, Sweden.
- Cui, A., Zhang, M., Liu, Y., & Ma, S. (2011). Are the URLs really popular in microblog messages?. In *2011 International Conference on Cloud Computing and Intelligence Systems* (pp. 1-5). IEEE, Beijing, China.
- Davidov, D., Tsur, O., & Rappoport, A. (2010). Enhanced sentiment learning using Twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 241-249). Association for Computational Linguistics, Stroudsburg, Pennsylvania, US.
- De Clercq, O., & Hoste, V. (2016). Rude waiter but mouthwatering pastries! An exploratory study into Dutch aspect-based sentiment analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (pp. 2910-2917). European Language Resources Association, Portorož, Slovenia.
- De Heer, W., & De Leeuw, E. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. *Survey nonresponse*, 41.
- Dhaoui, C., Webster, C. M., & Tan, L. P. (2017). Social media sentiment analysis: Lexicon versus machine learning. *Journal of Consumer Marketing*, 34(6), 480-488.
- Dubois, E., & Gaffney, D. (2014). The multiple facets of influence: Identifying political influentials and opinion leaders on Twitter. *Journal of American Behavioral Scientist*, 58(10), 1260-1277.
- Dwi Prasetyo, N., & Hauff, C. (2015). Twitter-based election prediction in the developing world. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media* (pp. 149-158). ACM, New York, US.
- Enli, G. (2017). Twitter as arena for the authentic outsider: Exploring the social media campaigns of Trump and Clinton in the 2016 US Presidential election. *European Journal of Communication*, 32(1), 50-61.
- Enli, G., & Skogerbø, E. (2013). Personalized campaigns in party-centred politics: Twitter and Facebook as arenas for political communication. *Journal of Information, Communication & Society*, 16(5), 757-774.
- Fanzilli, C. L. (2015). *Stock price prediction using sentiment detection of Twitter*. ORZO - Union College.
- Friese, M., Smith, C. T., Plischke, T., Bluemke, M., & Nosek, B. A. (2012). Do implicit attitudes predict actual voting behaviour particularly for undecided voters? *Journal of PLOS ONE*, 7(8), e44130.
- Garimella, K., Morales, G. D. F., Gionis, A., & Mathioudakis, M. (2018). Quantifying controversy on social media. *Journal of ACM Transactions on Social Computing*, 1(1), 1-26.
- Gautam, G., & Yadav, D. (2014). Sentiment analysis of Twitter data using machine learning approaches and semantic analysis. In *International Conference on Contemporary Computing* (pp. 437-442). IEEE, Noida, India.
- Gayo-Avello, D. (2012). "I wanted to predict elections with Twitter and all I got was this lousy paper": A balanced survey on election prediction using Twitter data. *arXiv preprint arXiv:1204.6441*.
- Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *Project Report, Stanford*, 1(12).

- Godin, F., Zuallaert, J., Vandersmissen, B., De Neve, W., & Van de Walle, R. (2014). Beating the bookmakers: Leveraging statistics and Twitter microposts for predicting soccer results. In *Proceedings of the Workshop on Large-Scale Sports Analytics* (pp. 1-4). ACM, New York, US.
- Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N., & Perera, A. (2012). Opinion mining and sentiment analysis on a Twitter data stream. In *International Conference on Advances in ICT for Emerging Regions* (pp. 182-188). IEEE, Colombo, Sri Lanka.
- Golbeck, J., Grimes, J. M., & Rogers, A. (2010). Twitter use by the US Congress. *Journal of the American Society for Information Science and Technology*, 61(8), 1612-1621.
- González, M. (2016). ¿Por qué el oficialismo tiene mayor presencia política en las redes sociales? [Why the officialism party has more presence in social media?]. *El Comercio*. Retrieved on March 29th, 2017 from <http://www.elcomercio.com/actualidad/oficialismo-elecciones-ecuador-redessociales-twitter.html>.
- González, M. (2017). La "guerra sucia" se libera en redes sociales ["Dirty war" is released on social media]. *El Comercio*. Retrieved on 29 March 2017 from <http://www.elcomercio.com/actualidad/guerra-sucia-ecuador-redessociales-elecciones.html>.
- Graham, T., Broersma, M., Hazelhoff, K., & Van'T Haar, G. (2013). Between broadcasting political messages and interacting with voters: The use of Twitter during the 2010 UK General election campaign. *Journal of Information, Communication, and Society*, 16(5), 692-716.
- Grant, W. J., Moon, B., & Busby Grant, J. (2010). Digital dialogue? Australian politicians' use of the social network tool Twitter. *Australian Journal of Political Science*, 45(4), 579-604.
- Haddi, E., Liu, X., & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Journal of Procedia Computer Science*, 17(2013), 26-35.
- Halper, D. (2016). How effective is the mainstream media at shaping public opinion? Retrieved on May 10th, 2016 from <https://www.quora.com/How-effective-is-the-mainstream-media-at-shaping-public-opinion>.
- He, S., Lee, S.-Y., & Rui, H. (2019). Open voice or private message? The hidden tug-of-war on social media customer service. In *Proceedings of the 52nd International Conference on System Sciences* (pp. 6638-6647). University of Hawaii at Manoa, Maui, Hawaii, US.
- Healy, D. (2015). The 2015 UK elections: Why 100% of the polls were wrong? *FTI Journal*. Retrieved on May 3rd, 2016 from <http://www.ftijournal.com/article/the-2015-uk-elections-why-100-of-the-polls-were-wrong>.
- Hern, A. (2015). Don't know the difference between emoji and emoticons? Let me explain. Retrieved on November 27th, 2018 from <https://www.theguardian.com/technology/2015/feb/06/difference-between-emoji-and-emoticons-explained>.
- Hoang, T. B. N., & Mothe, J. (2017). Predicting information diffusion on Twitter. Analysis of predictive features. *Journal of Computational Science*, 28(2017), 257-264.
- Hodges, D. (2015). Why did the polls get it wrong at the General election? Because they lied. Retrieved on May 3th, 2016 from <http://www.telegraph.co.uk/news/general-election-2015/politics->

blog/11695816/Why-did-the-polls-get-it-wrong-at-the-general-election-Because-they-lied.html.

- Hong, L., Doumith, A. S., & Davison, B. D. (2013). Co-factorization machines: Modelling user interests and predicting individual decisions in Twitter. In *Proceedings of the ACM International Conference on Web Search and Data Mining* (pp. 557-566). ACM, Rome, Italy.
- Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International Conference on Weblogs and Social Media* (pp. 216-226). AAAI Press, Ann Arbor, Michigan, US.
- Ilina, E. (2012). A user modeling oriented analysis of cultural backgrounds in microblogging. *Journal of Human Journal*, 1(4), 166-181.
- Jaidka, K., Ahmed, S., Skoric, M., & Hilbert, M. (2018). Predicting elections from social media: A three-country, three-method comparative study. *Asian Journal of Communication*, 29(3), 1-21.
- Jain, A., & Jain, V. (2019). Sentiment classification of Twitter data belonging to renewable energy using machine learning. *Journal of Information and Optimization Sciences*, 40(2), 521-533.
- Jain, V. (2013). Prediction of movie success using sentiment analysis of tweets. *The International Journal of Soft Computing and Software Engineering*, 3(3), 308-313.
- Jha, V., Manjunath, N., Shenoy, P. D., & Venugopal, K. (2016). Sentiment analysis in a resource scarce language: Hindi. *International Journal of Scientific and Engineering Research*, 7(9), 968-980.
- Jin, X., Gallagher, A., Cao, L., Luo, J., & Han, J. (2010). The wisdom of social multimedia: Using Flickr for prediction and forecast. In *Proceedings of the ACM Multimedia 2010 International Conference* (pp. 1235-1244). ACM, Firenze, Italy.
- Jose, R., & Chooralil, V. S. (2015). Prediction of election result by enhanced sentiment analysis on Twitter data using word sense disambiguation. In *2015 International Conference on Control Communication & Computing* (pp. 638-641). IEEE, Trivandrum, India.
- Joshi, P. A., Simon, G., & Murumkar, Y. P. (2018, August). Generation of brand/product reputation using Twitter data. In *International Conference on Information, Communication, Engineering and Technology* (pp. 1-4). IEEE, Pune, India.
- Jungherr, A., Jürgens, P., & Schoen, H. (2012). Why the pirate party won the German election of 2009, or the trouble with predictions: A response to Tumasjan, A., Sprenger, T.O., Sander, P.G., & Welpe, I.M. "Predicting elections with Twitter: What 140 characters reveal about political sentiment". *Journal of Social Science Computer Review*, 30(2), 229-234.
- Kachamas, P., Akkaradamrongrat, S., Sinthupinyo, S., & Chandrachai, A. (2019). Application of artificial intelligent in the prediction of consumer behavior from Facebook posts analysis. *International Journal of Machine Learning and Computing*, 9(1), 1-7
- Khan, F. H., Bashir, S., & Qamar, U. (2014). TOM: Twitter opinion mining framework using hybrid classification scheme. *Journal of Decision Support Systems*, 57, 245-257.
- Khan, M. L., Zaher, Z., & Gao, B. (2018). Communicating on Twitter for charity: Understanding the wall of kindness initiative in Afghanistan, Iran, and Pakistan. *International Journal of Communication*, 12, 25.

- Kharde, V., & Sonawane, P. (2016). Sentiment analysis of Twitter data: A survey of techniques. *arXiv preprint arXiv:1601.06971*.
- Kim, Y.-K., Lee, D., Lee, J., Lee, J.-H., & Straub, D. W. (2018). Influential users in social network services: The contingent value of connecting user status and brokerage. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, 49(1), 13-35.
- Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., & Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Journal of Nature Physics*, 6(11), 888-893.
- Kouloumpis, E., Wilson, T., & Moore, J. D. (2011). Twitter sentiment analysis: The good, the bad, and the OMG!. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (pp. 538-541). AAAI, Barcelona, Spain.
- Kumar, C. P., & Babu, L. D. (2019). Novel text pre-processing framework for sentiment analysis. *Smart Intelligent Computing and Applications* (pp. 309-317): Springer.
- Kumar, V., Park, H., Basole, R. C., Braunstein, M., Kahng, M., Chau, D. H., ... & Lesnick, B. (2014). Exploring clinical care processes using visual and data analytics: Challenges and opportunities. In *Proceedings of the 20th Conference on Knowledge Discovery and Data Mining Workshop on Data Science for Social Good* (pp. 1-5). ACM, New York, US.
- Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media?. In *Proceedings of the 19th International Conference on World Wide Web* (pp. 591-600). ACM, Raleigh, North Caroline, US.
- Kwakkel, J. H., Walker, W. E., & Haasnoot, M. (2016). Coping with the wickedness of public policy problems: Approaches for decision making under deep uncertainty. *Journal of Water Resources Planning and Management*, 142(3), 1-5.
- Larsson, A. O., & Ihlen, Ø. (2015). Birds of a feather flock together? Party leaders on Twitter during the 2013 Norwegian elections. *European Journal of Communication*, 30(6), 666-681.
- Larsson, A. O., & Moe, H. (2012). Studying political microblogging: Twitter users in the 2010 Swedish election campaign. *Journal of New Media & Society*, 14(5), 729-747.
- Le, H. T., Boynton, G., Mejova, Y., Shafiq, Z., & Srinivasan, P. (2017). Revisiting the American voter on Twitter. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 4507-4519). ACM, Denver, Colorado, US.
- Lee, J., & Lim, Y.-S. (2016). Gendered campaign tweets: The cases of Hillary Clinton and Donald Trump. *Journal of Public Relations Review*, 42(5), 849-855.
- Lee, J., & Xu, W. (2018). The more attacks, the more retweets: Trump's and Clinton's agenda setting on Twitter. *Journal of Public Relations Review*, 44(2), 201-213.
- Linh, D.-X., Stieglitz, S., Wladarsch, J., & Neuberger, C. (2013). An investigation of influentials and the role of sentiment in political communication on Twitter during election periods. *Journal of Information, Communication & Society*, 16(5), 795-825.
- Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. *Mining text data* (pp. 415-463): Springer.
- Loyola-González, O., López-Cuevas, A., Medina-Pérez, M. A., Camiña, B., Ramírez-Márquez, J. E., & Monroy, R. (2019). Fusing pattern discovery and visual

- analytics approaches in tweet propagation. *Journal of Information Fusion*, 46, 91-101.
- Maharani, W., Adiwijaya, & Gozali, A. A. (2014). Degree centrality and eigenvector centrality in Twitter. In *Eighth International Conference on Telecommunication Systems Services and Applications* (pp. 1-5). IEEE, Kuta, Indonesia.
- Mahmoudi, A., Yaakub, M. R., & Bakar, A. A. (2018). New time-based model to identify the influential users in online social networks. *Journal of Data Technologies and Applications*, 52(2), 278-290.
- Martínez-Cámara, E., Martín-Valdivia, M. T., Urena-López, L. A., & Montejo-Ráez, A. R. (2014). Sentiment analysis in Twitter. *Journal of Natural Language Engineering*, 20(1), 1-28.
- McCombs, M. (2014). *Setting the Agenda* (Second ed.): Cambridge: Polity Press.
- McShane, L., Pancer, E., & Poole, M. (2019). The influence of B to B social media message features on brand engagement: A fluency perspective. *Journal of Business-to-Business Marketing*, 26(1), 1-18.
- Meier, F., Elweiler, D. C., & Wilson, M. L. (2014). More than liking and bookmarking? Towards understanding Twitter favouriting behaviour. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media* (pp. 346-355). AAAI Press, Ann Arbor, Michigan, US.
- Mejova, Y., Srinivasan, P., & Boynton, B. (2013). GOP primary season on Twitter: Popular political sentiment in social media. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining* (pp. 517-526). ACM, New York, US.
- Mellon, J., & Prosser, C. (2017). Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Journal of Research & Politics*, 4(3), 1-9.
- Metaxas, P. T., Mustafaraj, E., Wong, K., Zeng, L., O'Keefe, M., & Finn, S. (2015). What do retweets indicate? Results from user survey and meta-review of research. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media* (pp. 658-661). AAAI Press, Oxford, UK.
- Metaxas, P. T., Mustafaraj, E., & Gayo-Avello, D. (2011). How (not) to predict elections. In *Third International Conference on Social Computing* (pp. 165-171). IEEE, Boston, Massachusetts, US.
- Miceli, M. (2015). Ecuador is Top on Twitter. *U.S. News*. Retrieved on 25 April 2017 from <https://www.usnews.com/news/articles/2015/07/17/ecuador-is-top-on-twitter>.
- Mitchell, A., & Hitlin, P. (2013). Twitter reaction to events often at odds with overall public opinion. Retrieved on November 15th, 2018 from <http://www.pewresearch.org/2013/03/04/twitter-reaction-to-events-often-at-odds-with-overall-public-opinion/>.
- Mukhtar, N., Khan, M. A., & Chiragh, N. (2018). Lexicon-based approach outperforms supervised machine learning approach for Urdu sentiment analysis in multiple domains. *Journal of Telematics and Informatics*, 35(8), 2173-2183.
- Myslín, M., Zhu, S.-H., Chapman, W., & Conway, M. (2013). Using Twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of Medical Internet Research*, 15(8), e174.
- Na, Y., & Kim, J. (2019). Sensibility and response keywords of users according to posting types of fashion Instagram focused on Koreans' fashion brands.

- Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., & Stoyanov, V. (2016). SemEval-2016 task 4: Sentiment analysis in Twitter. *arXiv preprint arXiv:1912.01973*.
- Nazir, F., Ghazanfar, M. A., Maqsood, M., Aadil, F., Rho, S., & Mehmood, I. (2019). Social media signal detection using tweets volume, hashtag, and sentiment analysis. *Journal of Multimedia Tools and Applications*, 78(3), 3553-3586.
- Neethu, M., & Rajasree, R. (2013). Sentiment analysis in Twitter using machine learning techniques. In *Fourth International Conference on Computing, Communications, and Networking Technologies* (pp. 1-5). IEEE, Tiruchengode, India.
- O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International AAAI Conference on Web and Social Media* (pp. 122-129). AAAI Press, Washington, US.
- Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Journal of Computers in Human Behavior*, 31, 527-541.
- Osmond, G. (2017). Tweet out? Twitter, archived data, and the social memory of out LGBT athletes. *Journal of Sport History*, 44(2), 322-335.
- Owen, D. (2017). New media and political campaigns. *The Oxford Handbook of Political Communication*.
- Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *International Journal of Advanced Research in Computer and Communication Engineering*, 10, 1320-1326.
- Pal, A., & Counts, S. (2011). Identifying topical authorities in microblogs. In *Proceedings of the ACM International Conference on Web Search and Data Mining* (pp. 45-54). ACM, New York, US.
- Pandarachalil, R., Sendhilkumar, S., & Mahalakshmi, G. (2015). Twitter sentiment analysis for large-scale data: An unsupervised approach. *Journal of Cognitive Computation*, 7(2), 254-265.
- Parackal, M., Mather, D., & Holdsworth, D. (2018). Value-based prediction of election results using natural language processing: A case of the New Zealand General election. *International Journal of Market Research*, 60(2), 156-168.
- Park, J., Na, Y., & Moon, I. C. (2017). Text augmented automatic statistician for predicting approval rates of politicians. In *2017 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 954-959). IEEE, Banff, Canada.
- Parmelee, J. H. (2014). The agenda-building function of political tweets. *Journal of New Media & Society*, 16(3), 434-450.
- Parmelee, J. H., & Bichard, S. L. (2012). *Politics and the Twitter revolution: How tweets influence the relationship between political leaders and the public*: Lexington Books.
- Parveen, H., & Pandey, S. (2016). Sentiment analysis on Twitter dataset using Naive Bayes algorithm. In *2016 International Conference on Applied and Theoretical Computing, and Communication Technology* (pp. 416-419). IEEE, Bangalore, India.
- Paul, M. J., Dredze, M., & Broniatowski, D. (2014). Twitter improves influenza forecasting. *Journal of PLoS Currents*, 6, 1-8.

- Peterka-Bonetta, J. (2015). Emoticons decoder for social media sentiment analysis in R. Retrieved on November 27th, 2018 from <https://github.com/today-is-a-good-day/emojis/>.
- R Core Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (Version 1.0.136). Retrieved from <http://www.r-project.org/>.
- Ramteke, J., Shah, S., Godhia, D., & Shaikh, A. (2016). Election result prediction using Twitter sentiment analysis. In *Proceedings of the International Conference on Inventive Computation Technologies* (pp. 1-5). IEEE, Tamilnadu, India.
- Rattananont, G., Toyoda, M., & Kitsuregawa, M. (2012). Characterizing topic-specific hashtag cascade in Twitter based on distributions of user influence. In *2012 Asia-Pacific Web Conference on Web Technologies and Applications* (pp. 735-742). Springer, Kunming, China.
- Rezapour, R., Wang, L., Abdar, O., & Diesner, J. (2017). Identifying the overlap between election result and candidates' ranking based on hashtag-enhanced, lexicon-based sentiment analysis. In *2017 11th International Conference on Semantic Computing* (pp. 93-96). IEEE, San Diego, US.
- Roccato, M., & Zogmaister, C. (2010). Predicting the vote through implicit and explicit attitudes: A field research. *Journal of Political Psychology*, 31(2), 249-274.
- Rofrío, D., Ruiz, A., Sosebee, E., Raza, Q., Bashir, A., Crandall, J., & Sandoval, R. (2019). Presidential elections in Ecuador: Bot presence in Twitter. In *2019 Sixth International Conference on eDemocracy & eGovernment* (pp. 218-223). IEEE, Quito, Ecuador.
- Rogers, E. M. (2010). *Diffusion of innovations* (4th edition ed.). New York: Simon and Schuster.
- Rui, H., Liu, Y., & Whinston, A. (2013). Whose and what chatter matters? The effect of tweets on movie sales. *Journal of Decision Support Systems*, 55(4), 863-870.
- Russell, F. M., Hendricks, M. A., Choi, H., & Stephens, E. C. (2015). Who sets the news agenda on Twitter? Journalists' posts during the 2013 US Government shutdown. *Journal of Digital Journalism*, 3(6), 925-943.
- Saif, H., He, Y., & Alani, H. (2012). Semantic sentiment analysis of Twitter. In *2012 International Conference on Semantic Web* (pp. 508-524). Springer, Boston, US.
- Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of Twitter. *Journal of Information Processing & Management*, 52(1), 5-19.
- Sang, E. T. K., & Bos, J. (2012). Predicting the 2011 Dutch Senate election results with Twitter. In *Proceedings of the Workshop on Semantic Analysis in Social Media* (pp. 53-60). Association for Computational Linguistics, Stroudsburg, Pennsylvania, US.
- Santos, J. C., & Matos, S. (2014). Analysing Twitter and web queries for flu trend prediction. *Journal of Theoretical Biology and Medical Modelling*, 11(1), S6.
- Schoen, H., Gayo-Avello, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M., & Gloor, P. (2013). The power of prediction with social media. *Journal of Internet Research*, 23(5), 528-543.
- Schumaker, R. P., Jarmoszko, A. T., & Labeledz Jr, C. S. (2016). Predicting wins and spread in the Premier League using a sentiment analysis of Twitter. *Journal of Decision Support Systems*, 88, 76-84.

- Schumaker, R. P., Solieman, O. K., & Chen, H. (2010a). *Predictive modeling for sports and gaming*: Springer.
- Schumaker, R. P., Solieman, O. K., & Chen, H. (2010b). Sports knowledge management and data mining. *Journal of Information Science and Technology Association*, 44(1), 115-157.
- Scott, J. (2017). *Social network analysis*: SAGE.
- Segesten, A. D., & Bossetta, M. (2017). A typology of political participation online: How citizens used Twitter to mobilize during the 2015 British General elections. *Journal of Information, Communication & Society*, 20(11), 1625-1643.
- Shafi, A., & Vultee, F. (2016). One of many tools to win the election: A study of Facebook posts by Presidential candidates in the 2012 election. *(R)evolutionizing political communication through social media* (pp. 210-228): IGI Global.
- Sharma, A., & Dey, S. (2012). A comparative study of feature selection and machine learning techniques for sentiment analysis. In *Proceedings of the 2012 ACM Research in Applied Computation Symposium* (pp. 1-7). ACM, San Antonio, Texas, US.
- Sharma, P., & Moh, T.-S. (2016). Prediction of Indian election using sentiment analysis on Hindi Twitter. In *International Conference on Big Data* (pp. 1966-1971). IEEE, Washington, US.
- Sharma, S., Rokne, J., & Alhadj, R. (2018). Predicting future with social media based on sentiment and quantitative analysis. *Applications of Data Management and Analysis* (pp. 199-209): Springer.
- Shattell, M., & Darmoc, R. (2017). Becoming a public thought leader in 140 characters or less: How nurses can use social media as a platform. *Journal of Psychosocial Nursing and Mental Health Services*, 55(6), 3-4.
- Sinnenberg, L., Buttenheim, A. M., Padrez, K., Mancheno, C., Ungar, L., & Merchant, R. M. (2017). Twitter as a tool for health research: A systematic review. *American Journal of Public Health*, 107(1), e1-e8.
- Skuz, M., & Romanowski, A. (2015). Sentiment analysis of Twitter data within Big data distributed environment for stock prediction. In *Federated Conference on Computer Science and Information Systems* (pp. 1349-1354). IEEE, Lodz, Poland.
- Smailović, J. (2014). *Sentiment analysis in streams of microblogging posts*. PhD thesis, Jožef Stefan International Postgraduate School, Ljubljana, Slovenia.
- Smailović, J., Grčar, M., Lavrač, N., & Žnidaršič, M. (2013). Predictive sentiment analysis of tweets: A stock market application. *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data* (pp. 77-88): Springer.
- Small, T. A. (2010). Canadian politics in 140 characters: Party politics in the Twitterverse. *Journal of Canadian Parliamentary Review*, 33(3), 39-45.
- Small, T. A. (2011). What the hashtag? A content analysis of Canadian politics on Twitter. *Journal of Information, Communication & Society*, 14(6), 872-895.
- Smith, M. A., Rainie, L., Shneiderman, B., & Himelboim, I. (2014). Mapping Twitter topic networks: From polarized crowds to community clusters. *Pew Research Center*, 20, 1-56.
- Socialbakers. (2017). Twitter statistics for Ecuador: Largest Audience. Retrieved on June 10th, 2017 from <https://www.socialbakers.com/statistics/twitter/profiles/ecuador/>.

- Stoessel, S. (2017, 10 June 2017). The left won Ecuador's presidential election. Cue right-wing revolt. *The Conversation*. Retrieved on December 12th, 2017 from <http://theconversation.com/the-left-won-ecuadors-presidential-election-cue-right-wing-revolt-76262>.
- Stoicescu, A. (2016). Hashtag(ing) between device and practice. *Analele Universitații București. Limba și literatura română*, 65, 95-108.
- Sweeney, C., & Padmanabhan, D. (2017). Multi-entity sentiment analysis using entity-level feature extraction and word embeddings approach. In *Proceedings of the International Conference Recent Advances in Natural Language Processing* (pp. 733-740). INCOMA Ltd, Varna, Bulgaria.
- Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Journal of Computational Linguistics*, 37(2), 267-307.
- Tang, L., Ni, Z., Xiong, H., & Zhu, H. (2015). Locating targets through mention in Twitter. *Journal of World Wide Web*, 18(4), 1019-1049.
- Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1), 163-173.
- Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12), 2544-2558.
- Tsakalidis, A., Papadopoulos, S., Cristea, A. I., & Kompatsiaris, Y. (2015). Predicting elections for multiple countries using Twitter and polls. *Journal of IEEE Intelligent Systems*, 30(2), 10-17.
- Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. *Journal of The International Linguistic Association*, 10(1), 178-185.
- Tunggawan, E., & Soelistio, Y. E. (2016). And the winner is...: Bayesian Twitter-based prediction on 2016 US Presidential election. In *2016 International Conference on Computer, Control, Informatics, and its Applications* (pp. 33-37). IEEE, Tangerang, Indonesia.
- Wang, L., & Gan, J. Q. (2018). Prediction of the 2017 French election based on Twitter data analysis using term weighting. In *Tenth International Conference on Computer Science and Electronic Engineering* (pp. 231-235). IEEE, Colchester, UK.
- Weng, J., Lim, E.-P., Jiang, J., & He, Q. (2010). TwitterRank: Finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (pp. 261-270). ACM, New York, US.
- Weng, J. H. (2019). Measuring the spreading of news on Twitter. Universidade Federal Do Rio Grande Do Sul
- Wickham, H. (2011). The Split-Apply-Combine strategy for data analysis. *Journal of Statistical Software*, 40(1), 1-29.
- Xu, K., Guo, X., Li, J., Lau, R. Y., & Liao, S. S. (2012). Discovering target groups in social networking sites: An effective method for maximizing joint influential power. *Journal of Electronic Commerce Research and Applications*, 11(4), 318-334.
- Vidal-Alaball, J., Fernandez-Luque, L., Marin-Gomez, F. X., & Ahmed, W. (2019). A new tool for public health opinion to give insight into telemedicine: Twitter poll analysis. *Journal of JMIR Formative Research*, 3(2), e13870.

- Zanfardini, M., Biasone, A. M., Bigné, E., & Ferri, K. A. (2015). ¿De qué hablan los turistas en la web? Estudio sobre el CGU de las marcas Valencia, Mar del Plata y San Martín de los Andes [What do tourists talk on the web? A study about UGC of the Valencia, Mar del Plata, and San Martín de los Andes brands]. In *Proceedings of the Fourth Latin American Congress of Tourism Research* (pp. 30-35). Quito, Ecuador.
- Zanfardini, M., Bigné, E., & Andreu, L. (2017). Análisis de la valencia y estrategia creativa del eWOM en destinos turísticos [Analysis of the valence and creative strategy of eWOM in tourism]. In *Proceedings of the 29th Marketing Congress* (pp. 1224-1236). ESIC, Sevilla, Spain.
- Zhao, J., & Gui, X. (2017). Comparison research on text pre-processing methods on Twitter sentiment analysis. *Journal of IEEE Access*, 5, 2870-2879.
- Zhou, T. (2011). Understanding online community user participation: A social influence perspective. *Internet Research: Electronic Networking Applications and Policy*, 21(1), 67-81.

Chapter 6

Predicting tweet impact using the evidential reasoning rule

Abstract: This study presents a novel evidential reasoning (ER) based prediction model called MAKER-RIMER, to examine how different features embedded on Twitter posts (tweets) may function as predictors of impact of tweets, in terms of the number of retweets achieved during an electoral campaign. The tweets posted by the two most voted candidates during the official campaign for the 2017 Ecuadorian Presidential election were used for this research. For each tweet, five features including type of tweet, emotion, URL, hashtag, and date are identified and coded to predict if tweets are either high or low impact. The main contributions of the new proposed model include its suitability to analyse tweet datasets based on likelihood data analysis. The model is interpretable, and the prediction process relies only on the use of available data. The experimental results show that MAKER-RIMER performed better, in terms of misclassification error, when compared against other predictive machine learning approaches. In addition, the model allows observing which features of the candidates' tweets are linked to high and low impact. Tweets containing allusions to the contender candidate, either with positive or negative connotations, without hashtags, and written towards the end of the campaign, were persistently those with the highest impact. URLs, on the other hand, is the only variable that performs differently for the two candidates in terms of achieving high impact. MAKER-RIMER can provide campaigners of political parties or candidates with a tool to measure how features of tweets are predictors of their impact, which can be useful to tailor Twitter content during electoral campaigns.

Keywords: Evidential reasoning rule, Machine learning, Twitter, Retweet, Prediction.

Declaration of interest: None.

6.1. Introduction

This paper develops a novel model based on the evidential reasoning (ER) rule for predicting the impact, high or low, that a tweet can achieve, in terms of its number of retweets. It does so by examining how various features embedded in tweets, namely type of tweet, emotion, uniform resource locator (URL), hashtag, and the moment in the timeline, relate to retweet counts during an electoral campaign. In this

paper, predictive model refers to a trained algorithm that classifies impact of tweets as high or low. The study uses the tweets produced by the two most voted candidates of the 2017 Ecuadorian Presidential election, Lenin Moreno and Guillermo Lasso, throughout the sixteen weeks of the official campaign period. The performance of the resulting ER based model is compared against other machine learning approaches used on Twitter analysis.

The ER based model proposed in this study, called MAKER-RIMER, comprises two approaches that combine techniques of likelihood data analysis and evidence-based probabilistic inference. MAKER stands for maximum likelihood evidential reasoning (Yang & Xu, 2017), and RIMER for belief rule-based inference methodology using the ER approach (Yang et al., 2007). The model aims at maximising the use of available data by splitting a model (MAKER) into sub-models (partial MAKER) for analysis and then combine them back together (RIMER).

MAKER-RIMER is meant to perform inference of data with uncertainty in an interpretable, transparent, and trackable way. In this context, uncertainty means a reduced ability of the algorithm to perform well, often due to incomplete knowledge (i.e. when no frequency exists for all possible combinations of parameters) (Kwakkel, Walker, & Haasnoot, 2016). On the other hand, “interpretable” means that the decisions made by the algorithm during the inference process are explicable to users (Laugel et al., 2018). “Transparent” implies that choices made during the model design process are visible to users (Varghese, Cawley, & Hong, 2018). And “trackable” denotes that the inference process can be tracked to determine how the input variables affect the output (Cheong & Gupta, 2005).

Concerning Twitter in electoral contexts, while the volume of users and information continues to expand globally, users and information continues to expand globally on Twitter, users find themselves in an increasingly contested environment when seeking to capture attention and spread their influence. This issue has gained particular relevance in electoral contexts. In the academic field, existing research has evidenced the critical role that Twitter plays in presidential elections. Indeed, there has been a dramatic increase in the use of Twitter for electoral purposes, and a progressive supplanting of traditional media platforms (Enli, 2017), especially when candidates feature limited political experience or lack support from influential actors from the political arena (Wang et al., 2016). This is evident at least for the period between Barack Obama’s victory in the 2008 US Presidential race (Cogburn &

Espinoza-Vasquez, 2011; Tumasjan, Sprenger, Sandner, & Welp, 2010), and the latest 2016 US Presidential election in which Donald Trump was elected president (Enli, 2017; Wells et al., 2016). Of relevance to this, several lines of inquiry have emerged that may contribute to identifying, for instance, who is reached on Twitter, who composes the intended audience, whether a message can have influence on political behaviour and preferences, and what impact a tweet can make.

However, although extensive research has been conducted on the role that Twitter plays during electoral campaigns, disagreements remain as an appropriate approach to measure the impact of a tweet. The most used metrics to address this issue are counts of account-followers, tweet favourites, and the number of retweets. Furthermore, less attention has been paid to what causes a tweet to be retweeted. Some studies suggest that the key to high retweet counts is to engage users with a high number of followers, often referred to as influential users, in coproducing content (De Veirman, Cauberghe, & Hudders, 2017; Keib, Himelboim, & Han, 2018). While this assertion is based on the belief that influential users have more visibility and power to influence a large number of users, scarce attention has been paid to what actually makes a tweet worth retweeting. Thus, the link between patterns in tweets and retweet counts remains an important subject of inquiry, particularly in connection with the claim that viral information reflects public opinions and political preferences (González-Bailón, Banchs, & Kaltenbrunner, 2012).

Following the work of Cha, Haddadi, Benevenuto, and Gummadi (2010), this paper assumes that the number of retweets embodies the influence of a tweet, and argues that retweets of a candidate's tweet are driven by a combination of content and name value. Therefore, the ability of a tweet to generate high impact is not a "one-size-fits-all" approach, but instead is mediated by both the candidate's profile and the content.

In the social media context, previous research has not fully addressed the mechanisms that underpin retweeting behaviour when deliberating about politics. When aiming to model Twitter data, previous studies have used machine learning methods such as logistic regression, decision tree, or support vector machine. A limitation of these methods is that they need "sufficiently large sample data to learn predictive models" (Kong et al., 2016, p. 36). However, even in the absence of statistically meaningful data to train a single model, these methods still proceed with

the prediction. It is likely, then, that their prediction outcomes might not be fully trusted when sufficient and meaningful data are not available.

This study seeks to address these issues by applying MAKER-RIMER on datasets that do not contain all value combinations of input variables or are not big enough. The sizes of datasets used for Twitter analysis vary widely across the relevant literature. To give a flavour, Soulier, Tamine, and Nguyen (2016) used 4.8 million tweets while Kavuluru and Sabbir (2016) used 1,000 tweets to train their models. While there is no consensus on what constitutes a big-enough dataset, there is agreement that traditional machine learning methods perform better when trained with large datasets (Rolnick, Veit, Belongie, & Shavit, 2017; Wong et al., 2020). Furthermore, the quantity of data has been associated to issues of data uncertainty (Beck, 1987). About this, Mahadevan and Sarkar (2009) have argued that uncertainty moderates as more information is obtained.

The foreseen advantages of using MAKER-RIMER in this study are twofold. First, it is likely that, given the number of tweets available for analysis and that the input variables do not have all value combinations, it performs better than other machine learning approaches as it is recursive in nature and can deal with incomplete datasets without deleting data or imputing data. This will be validated by comparison against other machine learning approaches. Second, the MAKER-RIMER model is purely data driven (Yang & Xu, 2017). This means two things: first, that it performs only with existing data, even when datasets are incomplete. The other machine learning methods, instead, often deal with incomplete datasets by relying on intuition, or by using data augmentation techniques (Wong et al., 2020), and second, that the weights that MAKER-RIMER would assign to the different parameters can show how the input variables influence the outcome, for which it is said to be a transparent model (Sabin, Xu, Chen, & Savan, 2013).

In addition to the MAKER-RIMER methodology, several machine learning approaches, namely logistic regression, Naïve Bayes, decision tree, and support vector machine, are also evaluated for prediction purposes to compare their performance based on misclassification errors (MCE) using the same datasets for training and testing purposes as in the MAKER-RIMER model. Lastly, the model described in this study can facilitate the identification of features of tweets that could lead to obtain a high number of retweets.

The rest of the paper takes the form of seven sections: In Section 6.2 the concepts of the ER rule are introduced, in which the MAKER and RIMER frameworks are presented. Section 6.3 reviews related work about predictive models using Twitter data. The methodology that leads to this study is described in Section 6.4, where the case study is introduced and the variables that are part of the model are presented. In Section 6.5 the case study using the MAKER-RIMER approach is conducted to predict the impact of tweets, as well as the application of different machine learning approaches for the same purpose. Finally, Section 6.6 shows the results and discussion, while the conclusion is presented in Section 6.7.

6.2. Brief introduction to the evidential reasoning rule

The evidential reasoning (ER) rule is based on the Dempster-Shafer (D-S) theory (Dempster, 1967; Shafer, 1976) and Bayesian probability theory. ER means reasoning with evidence (Srivastava, 2011). The ER rule is a probabilistic reasoning process to combine multiple pieces of independent evidence considering both reliability and weight of the evidence (Xu et al., 2020). A piece of evidence is independent if the information it contains does not depend on other evidence (Yang & Xu, 2013), and it is defined as a probability distribution over a set of mutually exclusive and collectively exhaustive propositions. Mutually exclusive means that propositions, which are the possible outcomes, cannot occur simultaneously. For this study, the outcome can be high or low impact, but not both. Collectively exhaustive, on the other hand, means that at least one of the possible events must occur. Again, the outcome must be only high or low.

Weight and reliability play an important role when considering the ER rule. Evidence weight, denoted by w_j , refers to the relative importance of the evidence, which can depend on the source and the way evidence is acquired (Yang & Xu, 2014). Evidence reliability, represented by r_j , denotes the ability of the information source to provide correct assessment to a problem (Smarandache, Dezert, & Tacnet, 2010). If all pieces of evidence, which are the observations obtained from the data, are acquired and measured in the same joint space, weight equals reliability, otherwise both need to be generated independently (Yang & Xu, 2014).

The ER rule consists of two parts: the bounded sum of the individual support of two pieces of independent evidence for each proposition, and the orthogonal sum of their collective support for each proposition, which makes it possible to combine

different pieces of evidence regardless of their order and without affecting the final results (Yang & Xu, 2013, 2014).

The ER rule has been applied in different disciplines and applications. For example, Zhu, Yang, Xu, and Xu (2016) have used ER to propose a model for monitoring asthma and manage its treatment in children, Xu et al. (2017) for data classification tasks across different kinds of database, and Fan, Yang, Perros, and Pei (2015) to identify trustworthiness in cloud computing services. Likewise, ER has been consistently applied in assessing navigational risk (Zhang, Yan, Zhang, Yang, & Wang, 2016), medical quality (Kong, Xu, Yang, & Ma, 2015), environmental impact (Wang, Yang, & Xu, 2006), and for conducting organization self-assessment (Xu & Yang, 2006). These examples suggest the versatility of ER for working with qualitative and quantitative data. The following subsections elaborate further on MAKER and RIMER frameworks.

6.2.1. The MAKER framework

The maximum likelihood evidential reasoning (MAKER), which is proposed by Yang and Xu (2017), is a methodological framework to combine multiple pieces of evidence under condition of uncertainty, such as randomness, inaccuracy, and ambiguity, for inferential modelling and analysis. Inferential modelling refers to the process of predicting outputs of a system from a set of inputs.

The MAKER framework demands the generation of joint frequency tables of the input variables, to then calculate basic probabilities or normalised likelihoods using Equation 6.1. In these calculations, the likelihood principle and the Bayesian principle need to be followed (Yang & Xu, 2014). When given two pieces of evidence $e_{i,l}$ and $e_{j,m}$ acquired from two variables x_l and x_m , at $x_l = x_{i,l}$ and $x_m = x_{j,m}$ respectively, their joint likelihood for proposition θ is represented by $c_{\theta,il,jm}$, which is the probability that both $x_{i,l}$ and $x_{j,m}$ are observed given proposition θ . Note that θ can be a single proposition or a subset of propositions. Then, the normalised likelihood is defined as follows (Yang & Xu, 2014, 2017)

$$p_{\theta,il,jm} = c_{\theta,il,jm} / \sum_{A \subseteq \Theta} c_{A,il,jm} \quad \forall \theta \subseteq \Theta \quad (6.1)$$

where $\Theta = \{h_1, h_2, \dots, h_N\}$ is defined as a frame of discernment and refers to a set of mutually exclusive and collectively exhaustive propositions.

Following the joint basic probability, the interdependence index is calculated to capture the statistical relationship between two pieces of evidence $e_{i,l}(A)$ and $e_{j,m}(B)$, and it is represented by $\alpha_{A,B,i,j}$. The interdependence index measures how strongly one input variable is related to another input variable. Since this index has been obtained from a space where basic probability is acquired as normalised likelihood, it needs to be scaled to ordinary likelihood (Yang & Xu, 2017). The formula to calculate the interdependence index is shown in Equation 6.2, while Equation 6.3 shows its properties

$$\alpha_{A,B,i,j} = \begin{cases} 0 & \text{if } p_{A,i,l} = 0 \text{ or } p_{B,j,m} = 0 \\ p_{A,B,il,jm} / (p_{A,i,l} p_{B,j,m}) & \text{otherwise} \end{cases} \quad (6.2)$$

$$\alpha_{A,B,i,j} = \begin{cases} 0 & \text{if } e_{i,l}(A) \text{ and } e_{j,m}(B) \text{ are disjoint} \\ 1 & \text{if } e_{i,l}(A) \text{ and } e_{j,m}(B) \text{ are independent} \end{cases} \quad (6.3)$$

After calculating the interdependence index, the next step is to generate the MAKER framework. In the MAKER framework, two pieces of evidence are combined to generate the combined support for proposition θ , as shown next. Suppose two pieces of evidence $e_{i,l}$ and $e_{j,m}$ are independent, the combined probability that proposition θ is jointly supported by both pieces of evidences is denoted by $p(\theta)$ as given by Equation 6.4

$$p(\theta) = \begin{cases} 0 & \theta = \emptyset \\ m_\theta / \sum_{c \subseteq \Theta} m_c & \theta \subseteq \Theta \end{cases} \quad (6.4)$$

where m_θ measures the combined probability mass for θ from both pieces of evidence and is generated as the bounded sum of the individual support for θ from both $e_{i,l}$ and $e_{j,m}$, and the orthogonal sum of their joint support with their interdependency and joint reliability taken into account, as shown in the recursive formula in Equation 6.5

$$m_{\theta} = [(1 - r_{j,m})m_{\theta,i,l} + (1 - r_{i,l})m_{\theta,j,m}] + \sum_{A \cap B = \theta} \gamma_{A,B,i,j} \alpha_{A,B,i,j} m_{A,i,l} m_{B,j,m} \quad (6.5)$$

where $r_{i,l}$ is the reliability of the evidence $e_{i,l}$. $\gamma_{A,B,i,j}$ is the ratio of the joint reliability over the product of the individual reliabilities of the two pieces of evidence $e_{i,l}$ and $e_{j,m}$ given that $e_{i,l}$ points to proposition A and $e_{j,m}$ to proposition B with $A \cap B = \theta$. Equation 6.5 should be first applied before Equation 6.4 is implemented.

6.2.2. The RIMER framework

RIMER is established as an extension of the traditional IF-THEN rules to beliefs rules (Yang et al., 2006). A belief rule is defined as a knowledge representation of information under uncertainty of vagueness or incompleteness (Chen et al., 2011; Zhang, Jiang, Chen, & Yang, 2015). In RIMER, an initial belief rule base (BRB) is constructed consisting in beliefs rules based on the knowledge of experts and experiences from users (Yang et al., 2006). Belief rule, denoted as R_k , is compounded of rule weights, antecedent attribute weights, and consequent belief degrees, and it is described as follows (Kong, Xu, Yang, & Ma, 2015)

$$R_k: \text{if } A_1^k \wedge A_2^k \wedge \cdots \wedge A_{T_k}^k, \\ \text{then } \{(D_1, \beta_{1k}), (D_2, \beta_{2k}), \cdots, (D_N, \beta_{Nk})\} \left(\beta_{jk} \geq 0, \sum_{j=1}^N \beta_{jk} \leq 1 \right), \quad (6.6)$$

with a rule weight θ_k , and attribute weights $\delta_1, \delta_2, \cdots, \delta_{T_k}$,
 $k \in \{1, \cdots, L\}$

where $A_i^k (i = 1, \cdots, T_k)$ is the referential category of the i^{th} antecedent attribute in the k^{th} rule, T_k is the number of antecedent attributes used in the k^{th} belief rule, $\beta_{jk} (j = 1, \cdots, N; k = 1, \cdots, L)$ is the assigned belief degree to consequent D_j which is used to describe input information that can be initially given by experts as subjective probability, $\delta_i (i = 1, \cdots, T_k)$ is the antecedent attribute weight that represents the relative importance of the i^{th} attribute, and θ_k is the rule weight representing the relative importance of the k^{th} rule. L represents the number of all belief rules in the rule base, and N is the number of all antecedent attributes used in the k^{th} rule.

The activation weight, denoted by w_k , is calculated for the k^{th} rule. The activation weight measures the degree to which the packet antecedent A^k in the k^{th} rule is activated by the input variables. The weight of each rule and degrees of belief should be considered. w_k is calculated as follows (Kong et al., 2015)

$$w_k = \frac{\theta_k \alpha_k}{\sum_{j=1}^L \theta_j \alpha_j} = \frac{\theta_k \prod_{i=1}^{T_k} (\alpha_{i,j}^k)^{\bar{\delta}_i}}{\sum_{l=1}^L [\theta_l \prod_{i=1}^{T_l} (\alpha_{i,j}^l)^{\bar{\delta}_i}]} \text{ and } \bar{\delta}_i = \frac{\delta_i}{\max_{i=1, \dots, T_k} \{\delta_i\}} \quad (6.7)$$

where $\theta_k (\in R^+, k = 1, \dots, L)$ is the relative weight of the k^{th} rule, and $\delta_i (\in R^+, i = 1, \dots, T_k)$ is the relative weight of the i^{th} antecedent attribute that is used in the k^{th} rule. The matching degree, $\alpha_{i,j}^k (i = 1, \dots, T_k)$, is the belief degree to which the input of the i^{th} antecedent attribute belongs to its j^{th} referential value $A_{i,j}^k$ in the k^{th} rule. This degree can be generated from different perspectives, depending on the nature and availability of the attributes (Yang, 2001; Yang et al., 2006). The final results are generated by aggregating all rules as described below

$$\mu = \left[\sum_{j=1}^N \prod_{k=1}^L \left(w_k \beta_{j,k} + 1 - w_k \sum_{i=1}^N \beta_{i,k} \right) - (N-1) \prod_{k=1}^L \left(1 - w_k \sum_{i=1}^N \beta_{i,k} \right) \right]^{-1} \quad (6.8)$$

where μ measures the degree to which the activation weight and belief degrees play in each rule.

$$\beta_j = \frac{\mu * [\prod_{k=1}^L (w_k \beta_{j,k} + 1 - w_k \sum_{i=1}^N \beta_{i,k}) - \prod_{k=1}^L (1 - w_k \sum_{i=1}^N \beta_{i,k})]}{1 - \mu * [\prod_{k=1}^L (1 - w_k)]}, j = 1, \dots, N \quad (6.9)$$

where β_j is a function of the belief degrees $\beta_{i,k} (i = 1, \dots, N, k = 1, \dots, L)$, the rule weights $\theta_k (k = 1, \dots, L)$, the attribute weights $\delta_i (i = 1, \dots, T)$, and the input vector x^* .

6.3. Related work

In its most abstract form, predictive models refer to the use of mathematical tools intending to predict future outcomes based on observed and assumed facts used as input variables (Chiu & Russell, 2011). Predicting an output includes, for example, foretelling future trends in behaviour patterns (Lin et al., 2012). Predictive models increasingly constitute a key supporting tool across a wide range of fields, such as marketing, health services, or fraud detection in the security systems industry. Nowadays, following the emergence and extensive use of social media platforms, vast amounts of data continuously generated and consumed by users, which contain valuable information about demographic aspects, preferences, and behaviours, are increasingly serving as grounds for predictive modelling (Bigsby, Ohlmann, & Zhao, 2019).

6.3.1. *Predictive models using Twitter data: Retweet analysis*

In recent years, there have been a growing number of publications focusing on predictive models based on Twitter, with the retweet measure being one of them. Retweeting refers to the act of sharing others' tweets within users' networks. The importance of the retweet lies in its ability to act as a dissemination tool, and to validate and engage with other Twitter users (Boyd, Golder, & Lotan, 2010). Retweet is equivalent to word-of-mouth (WOM) propagation in the Twitter context (Hochreiter & Waldhauser, 2013; Jin & Phua, 2014). It also serves as a metric used to determine the effectiveness, popularity, influence, and level of support of a given tweet or Twitter user account (Hong, Dan, & Davison, 2011; Nesi, Pantaleo, Paoli, & Zaza, 2018; Pezzoni et al., 2013; Suh, Hong, Pirolli, & Chi, 2010).

Studies about retweets have been conducted from different perspectives and with different approaches. For example, Lo, Chiong, and Cornforth (2016) focused on ranking audiences on Twitter, while Luo, Osborne, Tang, and Wang (2013) relied on the identification of retweeters, which are Twitter users that retweet others' tweets, to understand what prompts users to retweet. Rather than focusing on individual users, this approach demands focusing on the content that becomes widely shared. This perspective leads to the assumption that retweeting behaviour can be triggered by similarity of interests (Huang, Zhou, Mu, & Yang, 2014), as a reciprocity action (Berger & Milkman, 2012; Boyd et al., 2010; Lee, Kim, & Kim, 2015), to show support and agreement publicly (Boyd et al., 2010; Parmelee & Bichard, 2011), for

self-enhancement purposes to appear knowledgeable (Berger & Milkman, 2012; Lee et al., 2015), to build and engage in an online community (Kim, Sung, & Kang, 2014; Noriega, 2014; Zadeh & Sharda, 2014), for communication (Wang, Zuo, & Wang, 2015), and altruism purposes (Lee et al., 2015). Hence, understanding the motivations behind retweeting behaviour can be a complex task, but it is key when trying to connect with a target audience to disseminate content and gain influence.

Furthermore, another line of inquiry is concerned with what makes some tweets more likely to be retweeted than others. According to Pezzoni et al. (2013), the propensity of retweeting might be influenced by the position or visibility the tweets have, and by the number of followers Twitter user accounts have (Lee et al., 2015; Suh et al., 2010). The users with a high number of followers have greater probabilities of gaining a higher number of retweets than others (Bakshy, Hofman, Mason, & Watts, 2011; Kim et al., 2018). Yang et al. (2010) and Macskassy and Michelson (2011) also include posting time and sharing similar viewpoints in tweets as influential features for propagating tweets. Suh et al. (2010) state that, besides tweet content, the numbers of URLs and hashtags have a strong relationship with retweetability. Also, Savage, Monroy-Hernandez, and Höllerer (2016) observed that tweets containing call-to-action words, asking for the solidarity or empathy of users to act, e.g. “*Please RT to help...*”, have an influence on the propensity to retweet.

6.3.2. *Predicting retweets*

Retweets have been subject for testing different predictive models, some of which are based on the contents of topic-specific communities, and others on general content. Concerning models collecting random content, Petrovic, Osborne, and Lavrenko (2011) and Nesi et al. (2018) carried out studies to predict if tweets could be retweeted or not, without specific criteria when collecting tweets. From specific areas, on the other hand, retweeting behaviour studies have been conducted in fields like marketing (Kim et al., 2014), health (Kim, Hou, Han, & Himelboim, 2016), journalism (Trilling, Tolochko, & Burscher, 2017), and politics (Choi, 2014). Aiming to focus on an individual perspective, Xu and Yang (2012) and Choi (2014) agreed that tweet propagation is more affected by the way tweets are written than by their topic. A summary of these studies with their predictors is shown in Table 6.1. As is observed, there is not a unique or straightforward mechanism to analyse the propensity to retweet. Indeed, approaches seem to differ based on the context and

field of application. Although these models provide a deeper understanding of retweeting behaviour, this is mostly restricted to a number of features such as number of followers, URLs, hashtags, or mentions, for which there is a need to continue developing retweeting behaviour models by coding these features in a different way.

As shown above, developing predictive models of retweets have been of interest to a number of researchers in the last decade. In the political arena, retweets have been used as a proxy of vote share (Jaidka, Ahmed, Skoric, & Hilbert, 2018). Despite the ample literature focusing on retweets, however, unanswered questions remain. That is the case, for instance, of the name value of tweets. In other words, the disposition of Twitter user accounts to retweet content based on the author's fame or favouritism. Similarly, previous works have not addressed what works best for prominent politicians in terms of achieving retweets. These gaps have partially inspired the decision to model the retweets that politicians can achieve during electoral times, as presented in this study.

6.3.3. Contribution of this paper

Twitter data can yield effective and powerful indicators of future behaviour for a range of situations and applications. A primary concern of predictive models is still the capacity to deliver information in a dynamic fashion, which is useful to address opportunely the controllable features that affect the output variable. So far, predictive models have used metrics related to tweets or their authors in terms of numbers of URLs, hashtags, or followers to predict the retweeting behaviour. Meanwhile, there are features of tweets that remain unexplored, which can influence the propensity of users to retweet.

This study contributes to two areas of research: one is machine learning approaches to deal with data uncertainty. And the other, the use of Twitter data for prediction purposes. A combination of the MAKER and RIMER approaches based on the ER rule is developed, called MAKER-RIMER, to predict the impact of tweets based on the number of retweets, using only the available data. The main contribution of the proposed model is perhaps its interpretability. It means that the inference process is transparent and trackable during the application of the MAKER-RIMER model, and the results are interpretable in the sense that the outcomes are openly readable and understandable for users, whether they are academics or, in the context of this study, politicians or campaigners of a political party. This model involves the

use of codified features of tweets to analyse their influence on the number of retweets. In addition, the model allows the identification of the features of tweets that make the target audience more prone to retweet them. As a result of this study, a new approach to conduct machine learning when dealing with data uncertainty is offered, and a tool for campaigners of political parties to tailor and adapt tweets for enhanced retweetability.

Table 6.1

Summary of models aiming to predict retweeting behaviour

Authors	Aim	Field	Predictors	Approach
Petrovic et al. (2011)	Predicting retweetability of tweets. Data sources from experiment and tweets.	General topics	(1) Social features: number of followers, followees, favourites, lists, and (2) Characteristics related to tweet features: number of hashtags, mentions, URLs, length of tweets.	Passive-aggressive algorithm
Xu and Yang (2012)	Predicting retweetability of tweets from the perspective of specific users using Twitter as data source.	General topics	(1) Social-based showing the relationship between the tweet author and his/her network, (2) Content-based referring to the attractiveness of the tweet to draw attention from other users, (3) Tweet-based indicating the syntactic features of tweets, and (4) Author-based features representing the influence of the author of the tweet.	Support vector machine, logistic regression, and J48
Choi (2014)	Examining the role of emotions and cognitive processes in the frequency of retweeting posts using tweets as data source.	Politics	(1) Emotions, and (2) Cognitive processes.	Negative binomial regression
Kim et al. (2014)	Predicting retweetability of tweets using Twitter as data source.	Marketing	(1) Brand identification suggests that consumers identified with a brand tend to engage and to support activities related to the brand itself, (2) Brand trust refers to the reputation of the brand in terms of trustworthiness, (3) Community commitment is the level of social engagement between brand and customers, (4) Community membership intention implies customer's willingness to remain engaged to continue supporting the brand, and (5) Twitter usage frequency and number of tweets referring to customers engagement on Twitter activities.	Discriminant analysis
Kim et al. (2016)	Analysing the iteration and engagement on Twitter via retweet among cancer patients and their network. Only tweets with the keywords " <i>breast cancer</i> " were retrieved.	Health	(1) Structure of the social network, specifically number of followers, and (2) Content of the tweets, referring to the language and emotion used.	Logistic regression
Trilling et al. (2017)	Identifying the characteristics of news that are more likely to be retweeted, using data generated on Twitter and Facebook by news companies.	Journalism	(1) Geographical distance, (2) Cultural distance, (3) Negative content, (4) Positive content, (5) Conflict, (6) Human interest, and (7) Exclusiveness. In addition, control variables were included in the model, which referred to the topic of tweets, the number of days since the message was posted, and the length of the article.	Negative binomial regression
Nesi et al. (2018)	Predicting retweetability of tweets using online survey as main data source.	General topics	(1) Content of tweet: number of hashtags, mentions, URLs, and favourites; and publication time, (2) Author of the tweet: number of days since the Twitter user account was created, and number of tweets posted until date, and (3) Network: number of followers, followees, and listed count.	Principal component analysis

6.4. Methodology

This section presents the case study used in this research and provides details of the data sampling and analysis processes. The description of the model that constitutes the basis for the prediction of impact of tweets is also covered in this section.

6.4.1. Case study

The case study of the 2017 Ecuadorian Presidential election was conducted, in which the aim was to predict the impact of tweets in terms of number of retweets from the two most voted candidates, based on different features embedded in their own tweets.

6.4.2. Data sampling

All tweets produced by the two most voted candidates, Lenin Moreno from the ruling party, and Guillermo Lasso from the opposition, were extracted during the two rounds of the electoral campaign which lasted 16 weeks, by using the Twitter usernames of *@Lenin* and *@LassoGuillermo* respectively¹. As previously mentioned in Subsection 5.5.2 of Chapter 5, there was an additional Twitter user account obtaining a high number of retweets called *@VamosLenin* supporting Lenin Moreno, which would have been worth exploring for further analysis. However, this Twitter user account appeared as the most influential user in Week 4 as shown in Table 5.9, and at that point of time tweets generated during the first three weeks could not be retrieved. Nevertheless, the exclusion of this account does not influence the results of this study, which rather focuses on predicting if the tweets posted by the candidates themselves would achieve a high or low number of retweets.

Data extraction was performed by means of the Twitter Search API (application programming interface) and R Core Team (2013). Tweets produced by other Twitter user accounts which were retweeted by the candidates were removed from the datasets. Before starting the analysis of the data extracted, tweets were cleaned by replacing special Spanish accents using R. Four datasets were generated comprising the tweets each candidate generated during the first and second round of elections. These files also contained information about the number of retweets, which is the focus of this study, and other metrics such as date of creation, or number of

¹ At the data extraction time length of a tweet was limited to 140 characters.

favourites, also known as *likes*, which can be used to show appreciation for tweets². This comprised in total 650 tweets for Lenin Moreno and 1188 tweets for Guillermo Lasso, as shown in Table 6.2. This table also includes the number of retweets that candidates' tweets produced, descriptive statistics, and the number of followers and followees at the end of each round of voting. Although Guillermo Lasso posted almost twice as many tweets as Lenin Moreno, the latter candidate achieved more retweets. This might indicate that, in terms of drawing attention from users, Lenin Moreno's Twitter campaign was more successful, especially given that the number of followers Guillermo Lasso had was always higher than Lenin Moreno.

Table 6.2

Total number of tweets generated by the two candidates, total number of retweets generated by other Twitter users during the elections, and descriptive statistics of retweets

Candidates	1st round		2nd round	
	Lenin Moreno	Guillermo Lasso	Lenin Moreno	Guillermo Lasso
Total tweets	415	745	235	443
Total retweets	302,026	124,642	149,019	324,560
Min. retweets	13	9	6	54
Max. retweets	3,275	1,517	2,448	5,681
Median retweets	646	130	530	554
Mean retweets	727.77	167.30	634.12	732.64
No. followers	125,000	298,000	254,000	313,000
No. followees	25	1,456	26	1,457

In addition, 1.3 million tweets generated by Twitter users about the two candidates were collected to build the model as will be covered in Subsection 6.4.3.2. The criteria for extract the data were to retrieve tweets including mentions (*@Lenin*, *@LassoGuillermo*), hashtags (*#leninmoreno*, *#guillermolasso*), and keywords including candidates' names as previously described in Chapter 5, Subsection 5.4.3.

² <https://help.twitter.com/en/using-twitter/liking-tweets-and-moments>

6.4.3. Data analysis

With the assistance of R, tweets were randomly split for training and testing purposes. For each candidate, datasets from the first and second rounds were each split into five groups, 80% for training and 20% for testing purposes, accounting for a total of ten groups for each candidate. Thus, eight groups out of ten were used as training set (520 and 952 tweets for Lenin Moreno and Guillermo Lasso, respectively). And the remaining two groups for each candidate were used for testing purposes (130 and 236 tweets for Lenin Moreno and Guillermo Lasso, correspondingly).

In addition, the evolution of number of retweets was traced as shown in Figure 6.1. Retweets for Lenin Moreno's tweets had alternate peaks and troughs during the whole campaign, having the highest number of retweets during the seventh week which accounted for nearly 3,500 retweets. On the other hand, retweets for Guillermo Lasso's tweets increased progressively during the last week of the campaign, having the highest peak of retweets in the last week of the campaign with almost 6,000 retweets. There were eight candidates during the first round, which could have affected the proportion of retweets for Guillermo Lasso. However, since for the second round there were only two candidates, the number of retweets for Guillermo Lasso increased over time.

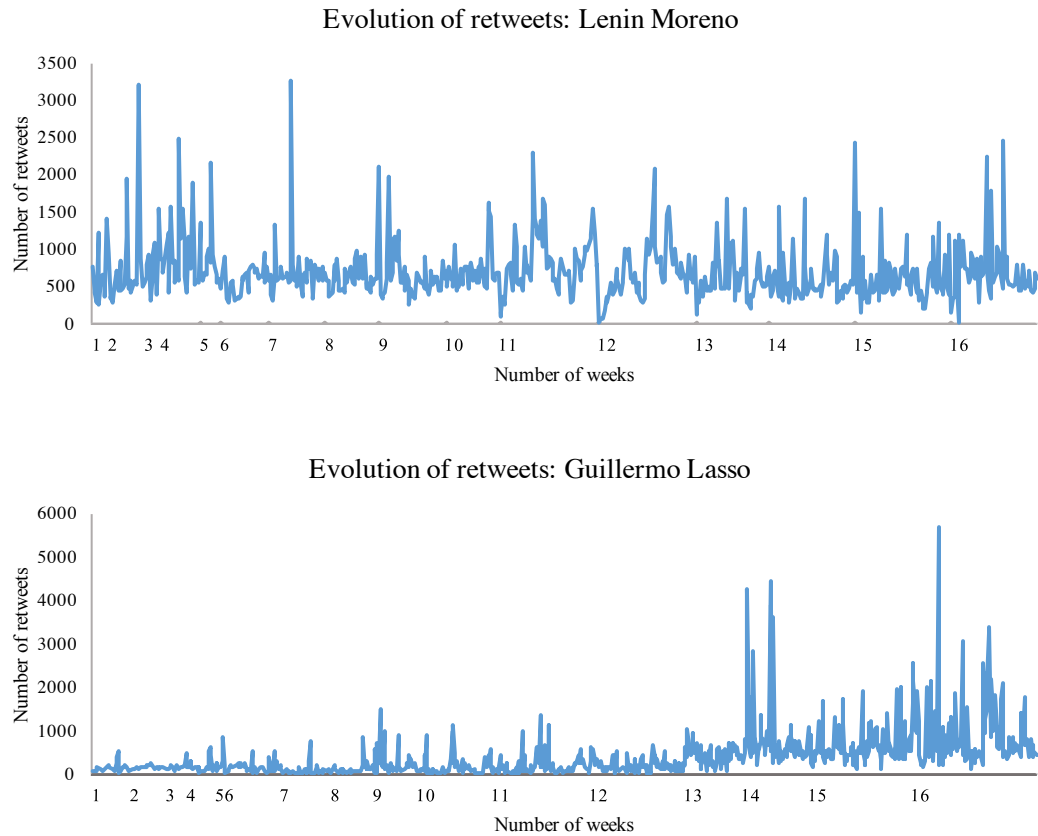


Figure 6.1. Evolution of the number of retweets for both candidates during the sixteen weeks of campaigning.

6.4.3.1. Description of the model: Output

The number of retweets achieved by each of the candidates' tweets is the metric used to measure the impact of tweets. The output of the model can take two categorical values, "high" or "low", which represents the impact of a tweet. For each candidate, a tweet is classified as high impact if the number of retweets it achieved is above the median of all the candidate's retweets achieved throughout the campaign. Otherwise, the tweet is labelled as low impact. The term high-impact tweet is in general used to refer to those tweets that reach a large number of users (Dabeer, Mehendale, Karnik, & Saroop, 2011). However, no criteria have been established for judging what constitutes a large number of retweets. Thus, a criterion emerging from the data needed to be defined for this research. A threshold was defined in statistical terms, as explained above, for each candidate, to satisfy the binary outcome required, high or low. Definition of this classification is consistent with previous works measuring the impact of Twitter in the academic field (Eysenbach, 2011; Thelwall,

Haustein, Larivière, & Sugimoto, 2013) and retweeting behaviour in high and low impact (Rudat & Buder, 2015). By using the median as a threshold, class balance is maintained, meaning that the outputs high and low are proportionally distributed across the datasets. In addition, the output of the model was purposely structured as a binary classification precisely for the integration of the MAKER-RIMER prediction. Also, the analysis used standard retweets. This means retweets of the original tweet as it is. Quote retweets, which are retweets that include a personal comment, were not included in the analysis because the Twitter Search API classifies them as independent tweets instead of conventional retweets. Finally, the Twitter Search API counts the number of retweets generated by the extended retweeting network, which means that a retweet can subsequently be retweeted by other users.

6.4.3.2. Description of the model: Inputs

André, Bernstein, and Luther (2012) claim that the ideas contained in tweets may influence the propensity of retweeting. This suggests that, content of tweets can help attract new followers and engage them over time (Araujo, Neijens, & Vliegenthart, 2015; Tan, Lee, & Pang, 2014). This study investigates retweets in terms of the ways and moment in which politicians post tweets. Therefore, the results are of particular relevance to electoral campaigns. For this purpose, the information about the features of the tweets embedded in the candidates' tweets is extracted and classified into five variables as shown in Figure 6.2, which constitute the input of the model, as described below:

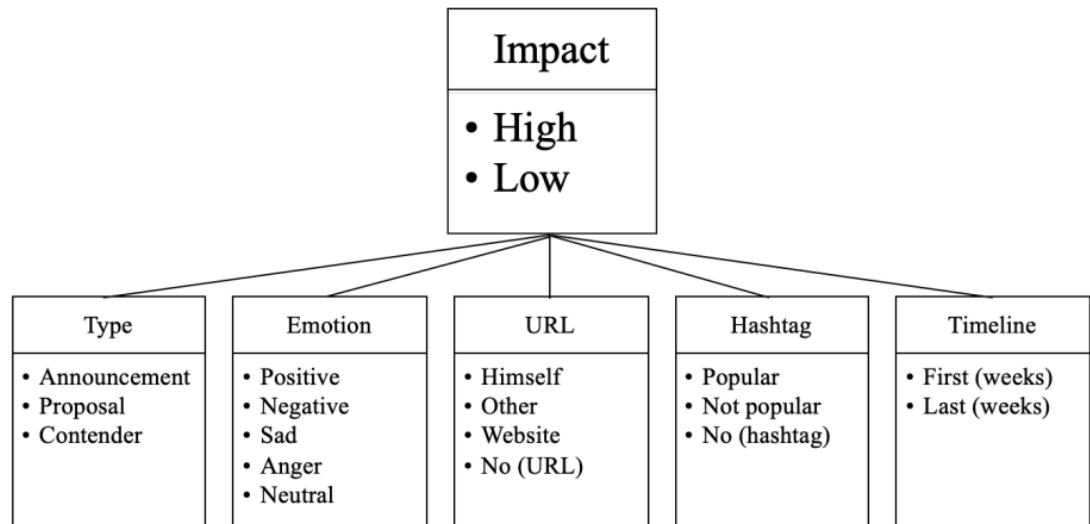


Figure 6.2. Original model developed to predict the impact of a tweet based on the number of retweets.

Input variables were selected on the basis of previous studies which suggest that:

- Politicians strive to communicate directly with voters and stay relevant, for which Twitter is a convenient platform, particularly during electoral times (Fountaine, 2017; Lee & Xu, 2018). To position themselves, politicians need to generate news continuously (Weingart, 2007). So, choosing among different types of messages for posting is a recurring task (Bode & Vraga, 2018). In this study, the candidates Lenin Moreno and Guillermo Lasso posted on average six and ten tweets per day respectively, with different purposes, throughout the campaign.
- Politicians use emotional communication to transfer feelings (Entman, 1992). They would often communicate “contentious issues in partisan terms” (Fogarty & Wolak, 2009, p. 134), so it is expected that their tweets have emotional content.
- URLs are popular on Twitter for communicating ideas by redirecting to a video, audio or image hosting platform (Maity, Gajula, & Mukherjee, 2018). This study reflects such popularity, since Lenin Moreno and Guillermo Lasso used 547 and 797 URLs in their tweets respectively during the campaign. That is, correspondingly, 84% and 67% of their tweets included an URL.
- Hashtags are also popular on Twitter for propagating ideas and promoting topics (Cunha et al., 2011). In this study, Guillermo Lasso used hashtags in 532 tweets (45% of them) while Lenin Moreno in only 67 (10% of them). This difference in

the intensity of hashtag-use offered the opportunity to explore how hashtags relate to the number of retweets.

- Finally, previous studies have concluded that the time tweets are written can affect the retweeting behaviour (Feng & Wang, 2013; Lee & Xu, 2018). Similarly, it is expected that polarisation increases as the campaign goes forward, as well as the attention to candidates. The following paragraphs provide details of how the input variables are coded for this study.

Type of tweet: This variable represents the type of messages that the candidates posted during the campaign in terms of its purpose. Ramos-Serrano, Fernandez-Gomez, and Pineda (2018) identified that in the 2014 European Parliament election in Spain, most candidates used Twitter for unidirectional communication, and very few for interactivity with other users. When used unidirectionally, candidates posted tweets mainly for self-promotion and to supplement their offline interactions (Lim, 2018). In addition, Ott (2017) observed that Donald Trump used Twitter to embarrass or criticise his opponents in the 2016 US Presidential election. By looking at a random sample of 800 tweets by Lenin Moreno and Guillermo Lasso, almost all of them were unidirectional, meaning that they did not reply to other Twitter users. Instead, candidates' tweets either mentioned the opponent, a campaign issue or topic, or were used for announcing events. Thereafter, tweets are manually analysed and categorised into three groups: "contender", "proposal", or "announcement". A tweet is classified as "contender" if it contains any type of information about the other candidate through hashtags, mentions, names or any other reference, for example, the following tweet written by Guillermo Lasso: *"It is comprehensible that @Lenin does not know how to create jobs because he has never done so. I have experience in the private sector"*. If a tweet has information about their own campaign proposals, they are classified as "proposal", for example *"I will derogate the Communication Law"*. Finally, if a tweet does not contain information about their agendas or topics that are considered either contender or proposal, it is labelled as "announcement", and this type of tweet could include tweets such as *"Good morning! We are starting the interview with @desayunos24 in @teleamazonasec. Don't miss it!"*.

Emotion: The next step is to categorise tweets based on emotion, which is known as emotion analysis. Emotion analysis aims to detect moods based on a specific text, such as happiness, sadness, fear, anger, disgust, and surprise (Ekman,

1993). Emotion differs from sentiment analysis in that emotion relates to people's mood and is determined by a multi-class classifier that includes, for instance, anger, fear, and surprise. Sentiment, on the other hand, is associated with users' feelings and opinions, and is usually measured using a binary classification of positive and negative (Allouch, 2018; Kaur & Saini, 2014). For this classification, a text analytic software tool, the Linguistic Inquiry and Word Count software (LIWC2015) (Pennebaker, Boyd, Jordan, & Blackburn, 2015), is used to analyse the emotions of tweets. By using LIWC2015, the following emotions were detected in tweets: "positive", including words such as love, happy, and nice; "negative", containing words such as hurt, ugly, and nasty; "sadness", embracing words such as crying, grief, and sad; "anger", including words such as hate, kill, annoyed, and pissed. Finally, if LIWC2015 cannot detect any emotion, it is manually labelled as "neutral".

URL: Due to Twitter's character limitation, the use of a URL as a link to an external website can offer deeper content for other Twitter users. When the tweets are extracted from Twitter using the Search API, information such as images, videos, or GIFs are also converted into internal Twitter URLs. Concerning categories in which to classify URLs, this is an issue that the literature has not addressed thoroughly. Apparently, the criteria for categorisation depend on researchers' judgements. Studies concerned with the detection of harmful or malicious content, for instance, have typically used a binary approach to categorisation such as suspect/normal (Hammami, Chahir, & Chen, 2003), or benign/malicious (Nagaonkar & Kulkarni, 2016). To make classification relevant for this study, URLs were manually analysed. That is, each URL was visited and labelled in terms of: a) the type of file to which the URL redirects. Maharana, Nayak, and Sahu (2006) developed a categorisation of URL into six file formats: html, pdf, ppt, doc, rtf, and others. In this work, however, URLs most often redirected to a website, a video, an image, or a GIF. And b) the topic or subject associated to the URL (Zandona, Rault, & Ripsher, 2013). In most cases, the URLs retrieved would redirect to image or video files for self-promotion of the candidates themselves. This evidence helped to categorise URLs into: "himself", "other", "website", or "no". "Himself" is assigned if the internal URL contains information about the candidate under analysis, such as images or videos; "other", if the internal URL contains information about other people or situations not directly related to candidate under analysis; "website" if the URL is redirected to an external website; or "no", meaning that a tweet does not contain URLs.

Hashtag: On Twitter, a hashtag refers to a word, or a phrase preceded by the hash sign “#”, which is used as a keyword to identify specific topics. Hashtags were classified into three categories: “popular”, “not popular”, and “no”. This study assumes that candidates used hashtags to gain awareness from others (Hemphill, Culotta, & Heston, 2013). So, the popularity of hashtags is relevant for that purpose. Popularity was measured using a ranking procedure. Hashtags were extracted from the 1.3 million tweets dataset used in Subsection 5.4.3 of Chapter 5 and arranged in a spreadsheet document in terms of their frequency of occurrence in the dataset, from high to low. When the hashtags used in the candidates’ tweets appeared among the twenty most popular hashtags of the ranking, the tweet is labelled as “popular”, and otherwise “not popular”. The twenty most popular hashtags covered almost 80% of the hashtag-presence in the 1.3 million tweets dataset, as shown in Figure 6.3 and Figure 6.4. In these figures, the x- and y-axes change depending on the number of hashtags found in the datasets of each candidate, and the frequency with which the hashtags were shared. If a tweet does not have the presence of hashtags, “no” is assigned to this variable. Also, if a tweet contains more than one hashtag, all the hashtags are tested for popularity according to the procedure explained above. When at least one of the hashtags appeared among the twenty most popular hashtags of the ranking, it is labelled as “popular”. In addition, a weekly breakdown of the frequency of hashtags is presented in Figure A-6.1, Figure A-6.2, Figure A-6.3, and Figure A-6.4 in the Appendix A-6. Figure A-6.1 and Figure A-6.2 show the frequency of hashtags per week for Lenin Moreno and Guillermo Lasso respectively throughout the campaign, as well as the frequency of the most shared hashtags. It can be observed that the number of hashtags tends to increase for both candidates as the campaign goes forward and reaches their picks in week 15. The popularity of most-shared hashtags is different for each candidate and changes every week, for which the scales of the y-axes differ in the plots of Figure A-6.3 and Figure A-6.4.

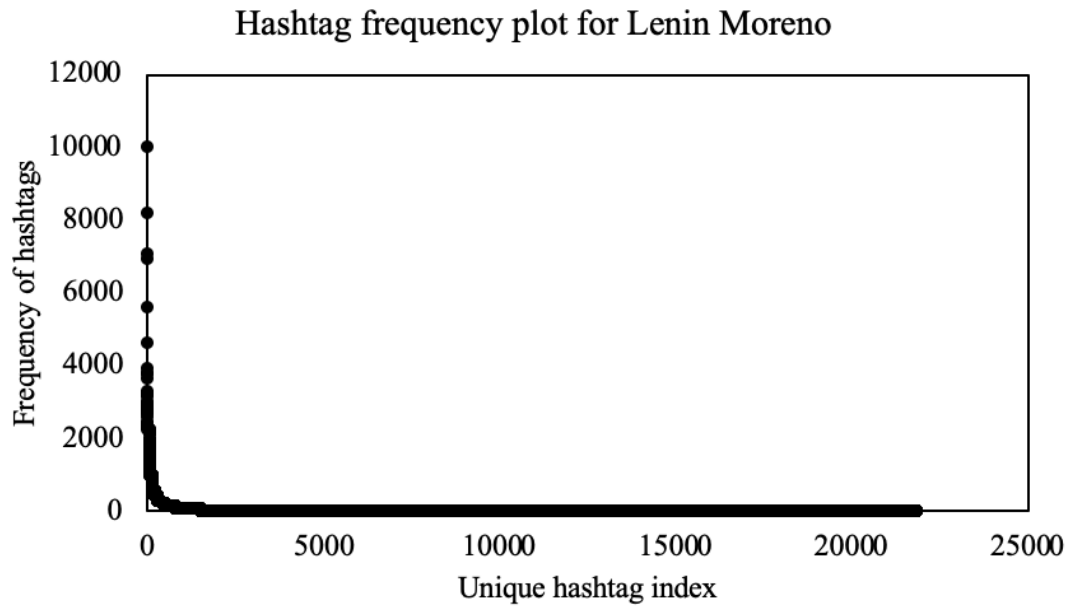


Figure 6.3. Frequency plot of hashtags found in tweets about Lenin Moreno.

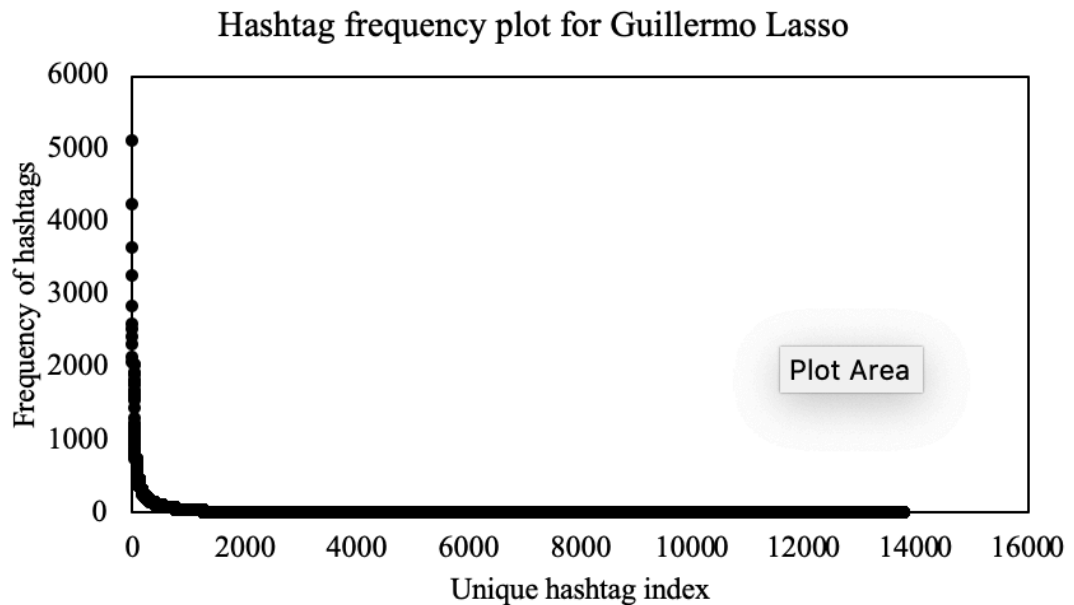


Figure 6.4. Frequency plot of hashtags found in tweets about Guillermo Lasso.

Timeline: This variable reflects the time when the tweet was posted. Since the campaign lasted for 16 weeks, “first” is assigned if tweets are written during the first eight weeks of the campaign; otherwise, they are labelled as “last”. There is no clear set of rules in previous research regarding the decision to split a timeline for modelling. When the timeline is of binary type, a “before” and “after” approach has been used widely (Muggler, Eshwarappa, & Cankaya, 2017; Tamaddoni, Stakhovych,

& Ewing, 2017; Zhong et al., 2016). In this study, a seemingly natural way to split the timeline is first and second round, that is, weeks 12 and 16 respectively. However, the data show an increasing trend of tweets posted by both candidates every week as shown in Figure 6.5, while the median weekly tweet-count was 38 and 67 for Lenin Moreno and Guillermo Lasso, respectively. Except for week 7 for Lenin Moreno's campaign and week 6 for that of Guillermo Lasso, all the values (counts of tweets posted by each candidate) below the median occurred during the first eight weeks, instead of the first twelve weeks. Thus, the first and last eight weeks are two different temporal contexts, for which the timeline was split at that point.

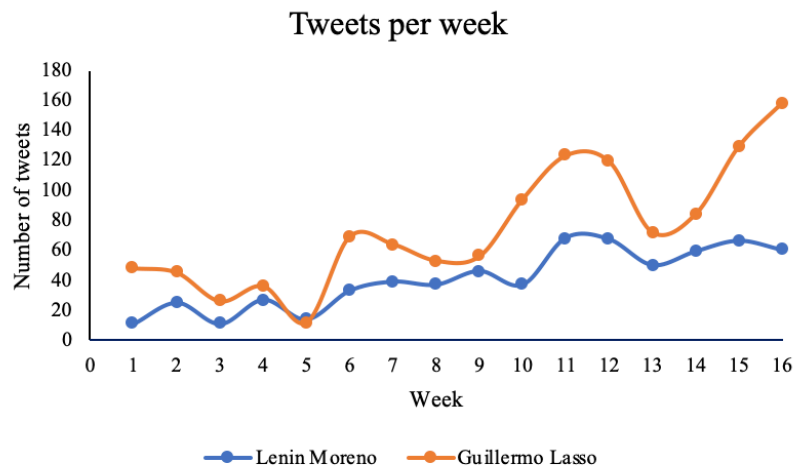


Figure 6.5. Candidates' tweets per week.

6.5. Application of the ER rule to predict impact of tweets based on the number of retweets for the 2017 Ecuadorian Presidential election

This section presents the MAKER-RIMER prediction model, which is based on the ER rule, used in this study. The prediction model aims to determine the impact of the tweets posted by the two most voted candidates of the 2017 Ecuadorian Presidential election, Lenin Moreno and Guillermo Lasso, in terms of the number of retweets their tweets achieved. The following subsections deal with the implementation of MAKER, the implementation of RIMER, the method used to train the different parameters, and the comparison of the model against other machine learning approaches.

6.5.1. Implementing the MAKER framework

If the datasets contained all value combination of input variables, meaning a situation in which frequencies exist for all possible combination of the parameters introduced in Figure 6.2, a single model with all five input variables could be used to develop a predictive model for each one of the two candidates. In this study, however, developing such a single model could lead to misleading results because of the absence of sufficient data for taking into account the five variables together (McDonald, 2014; Yates, 1999). Since the datasets for this study do not contain all possible combinations of input variables, two partial MAKER models for each candidate, needed to be trained to generate the prediction of the impact of tweets. The combination of variables in each partial MAKER model needs to have sufficient data to calculate the joint probabilities. Each partial MAKER model comprises two and three variables which need to be closely correlated to each other to calculate joint probabilities. The partial models are shown in Figure 6.6 in Subsection 6.5.2. In both partial MAKER models, the output would be the same which is high or low impact.

To select the combination of variables to be grouped into each partial MAKER model, the rationale is to choose the combination of variables with the highest number of records evenly distributed in the space model which shows the relationship between the input and output variables (Yang & Xu, 2017). This is done by registering the frequencies for all possible combinations of two and three parameters of the input variables presented in Figure 6.2. Using two partial MAKER models instead of a single one is justified in that it allows increasing the use of the available data. For example, it is likely that in a single model, no frequency exists for a combination of parameters *announcement*, *sad*, *other*, *not popular*, and *first*. Inversely, if splitting a model into two partial models of inputs Type/Emotion, and of URL/Hashtag/Timeline, it is likely that frequencies exist for combinations of *announcement* and *sad*, and for *other*, *not popular*, and *first*. For this study, some combinations of variables do not have records available if a single model were to be developed.

From each combination of input variables selected, joint probability tables are generated. This is done by recording the frequencies for the inputs and outputs variables from the observations. Then, the interdependent index is calculated, which shows the statistical interrelationship between each group of variables. From these results, the partial MAKER model is trained to predict the impact of the tweets. The

weights of the five variables and their parameters are trained for optimal prediction. Optimal prediction refers to maximising the likelihood of true state being generated by using the model, while minimising the errors between the output of the observation and the predicted one. The parameters of the variables are the sub-categories of each input variable in the partial MAKER models. For instance, the input variable “type” can have three possible parameters namely “announcement”, “proposal” or “contender” as previously shown in Figure 6.2. An example using the data from the first partial MAKER model from Guillermo Lasso, comprising the input variables type and emotion, is following presented in Table 6.3, Table 6.4, and Table 6.5.

After defining the partial MAKER models, which in the example of Guillermo Lasso involves input variables type and emotion, the first step in the implementation of MAKER is the creation of the joint frequency tables (i.e. columns 3 and 4 in Table 6.3) while the result of the output variable can be high or low impact. The estimates likelihood that the pieces of evidence of the input variables are high or low are presented in $c_{\theta,il,jm}$ as shown in columns 5 and 6 from Table 6.3. Then, the normalised likelihoods are calculated as shown in columns 7 and 8 in Table 6.3 using Equation 6.1 to estimate the joint basic probability. For the partial MAKER models presented in this case study, the pieces of evidence are the observations of the parameters of the input variables and the output. For example, the pieces of evidence of the first partial MAKER model for Guillermo Lasso comprise a parameter for emotion (positive, negative, sadness, anger, neutral), a parameter of type of tweet (announcement, proposal, contender), and an impact for the tweet (high or low). Each piece of evidence is represented as an extended probability distribution or belief distribution, with probabilities assigned to propositions, e.g. singleton propositions such as high and low, or non-singleton propositions such as the set of high or low. This way, ambiguity (or unknown) caused by missing data can be represented as probabilities assigned to non-singleton propositions such as the set of high or low.

Table 6.3

Estimates and normalised likelihoods for the first partial MAKER model of Guillermo Lasso comprising the variables type and emotion

Input variables		Frequencies		Estimates likelihood $c_{\theta,il,jm}$		Normalised likelihood $p_{\theta,il,jm}$	
		Output Impact		Output Impact		Output Impact	
		High	Low	High	Low	High	Low
Positive	Announcement	142	109	0.2971	0.2300	0.5637	0.4363
	Proposal	241	240	0.5042	0.5063	0.4989	0.5011
	Contender	15	1	0.0314	0.0021	0.9370	0.0630
Negative	Announcement	8	5	0.0167	0.0105	0.6134	0.3866
	Proposal	5	0	0.0105	0.0000	1.0000	0.0000
	Contender	13	0	0.0272	0.0000	1.0000	0.0000
Sadness	Announcement	1	7	0.0021	0.0148	0.1241	0.8759
	Proposal	0	2	0.0000	0.0042	0.0000	1.0000
	Contender ³	0	0	0.0000	0.0000	0.0000	0.0000
Anger	Announcement	9	4	0.0188	0.0084	0.6905	0.3095
	Proposal	4	1	0.0084	0.0021	0.7987	0.2013
	Contender	7	0	0.0146	0.0000	1.0000	0.0000
Neutral	Announcement	25	91	0.0523	0.1920	0.2141	0.7859
	Proposal	5	13	0.0105	0.0274	0.2761	0.7239
	Contender	3	1	0.0063	0.0021	0.7484	0.2516

After obtaining the joint basic probability, the interdependence index is calculated between each pair of evidence and it is presented in columns 3 and 4 of Table 6.4. In addition, normalised likelihood is scaled to ordinary likelihood as shown in columns 5 and 6 in the same table using Equation 6.2.

³ For this combination of parameters, data were not available for calculating joint probabilities and prediction.

Table 6.4

Interdependence index applied for the first partial MAKER model of Guillermo Lasso

Input variables		Interdependence index			
		Normalised likelihood		Ordinary likelihood	
		Output		Output	
		High	Low	High	Low
Positive	Announcement	2.3158	1.7168	5.8175	4.3128
	Proposal	1.8945	2.1191	3.1620	3.5369
	Contender	1.8618	2.6593	7.3116	10.4433
Negative	Announcement	1.5946	4.4016	3.1985	8.8288
	Proposal	2.4027	0.0000	16.0140	0.0000
	Contender	1.2573	0.0000	0.2513	0.0000
Sadness	Announcement	2.7223	1.7983	2.8683	1.8948
	Proposal	0.0000	2.2068	0.0000	11.8356
	Contender	0.0000	0.0000	0.0000	0.0000
Anger	Announcement	1.8826	2.8425	3.0482	4.6025
	Proposal	2.0124	1.9878	10.8021	10.6699
	Contender	1.3186	0.0000	0.3949	0.0000
Neutral	Announcement	1.9620	1.9063	1.9666	1.9108
	Proposal	2.3384	1.8874	19.2526	15.5395
	Contender	3.3170	6.5472	9.6216	18.9914

After calculating the interdependence index, the next step is to generate the partial MAKER models using Equation 6.5 and Equation 6.4. For Guillermo Lasso, the first partial MAKER model comprises the input variables: type and emotion, while for the second partial MAKER model, the input variables are URL, hashtag, and timeline as seen in Figure 6.6. The model assumes that initial weights are the same for all the input variables and their parameters, but these weights are later trained or optimised as will be explained in Subsection 6.5.3. In addition, since the data come from the same source, weight is equal to reliability in this study as previously introduced in Section 6.2. In the MAKER framework, two pieces of evidence are combined to generate the combined support for proposition θ , which for this case study refers to the output variable being high or low, and it is defined by using Equation 6.5 and Equation 6.4, and the results are shown in Table 6.5. This table presents the results of the first partial MAKER model after training the weights of the parameters of MAKER. The table shows, for Guillermo Lasso, the probabilities of a tweet achieving high or low impact for each combination of inputs type and emotion. For example, a type of tweet coded as “announcement” with an emotion

coded as “positive”, has a probability of 0.5877 of being high impact, and 0.4123 of being low impact.

Table 6.5

First partial MAKER model results after training weights of variables (type and emotion) and their parameters for Guillermo Lasso

Input variables		MAKER 1 results after training weights	
		Output	
		High	Low
Positive	Announcement	0.5877	0.4123
	Proposal	0.5026	0.4974
	Contender	0.9741	0.0259
Negative	Announcement	0.6280	0.3720
	Proposal	0.9484	0.0516
	Contender	0.9035	0.0965
Sadness	Announcement	0.2224	0.7776
	Proposal	0.0672	0.9328
	Contender	0.0000	0.0000
Anger	Announcement	0.6967	0.3033
	Proposal	0.8656	0.1344
	Contender	0.9253	0.0747
Neutral	Announcement	0.2601	0.7399
	Proposal	0.1807	0.8193
	Contender	0.8013	0.1987

This process is repeated with the second group of variables that form the second partial MAKER model, which for Guillermo Lasso comprises the three following input variables: URL, hashtag, and timeline (Table A-6.4 in the Appendix A-6). These are the three variables left after the first partial MAKER model was generated, which complies with the condition of variables being closely correlated to calculate joint probabilities. A similar calculation process was conducted for the two partial MAKER models for Lenin Moreno, and the results are presented in Table A-6.2 and Table A-6.3 in Appendix A-6.

6.5.2. Implementing the RIMER framework

After completing the partial MAKER models for each candidate, a RIMER model is developed to combine the results generated by the two partial MAKER

models, as shown in Figure 6.6, where MAKER 1 is the first partial MAKER model comprising two input variables, and MAKER 2 is the second partial MAKER model comprising three input variables for each candidate. Thus, RIMER combines the two partial MAKER models back together. As previously explained in Subsection 6.5.1, for each candidate the input variables for the two partial MAKER models were grouped together only if the groups of variables were closely correlated to each other based on the available data for each candidate. That is the reason why partial MAKER models for each candidate are composed of different input variables.

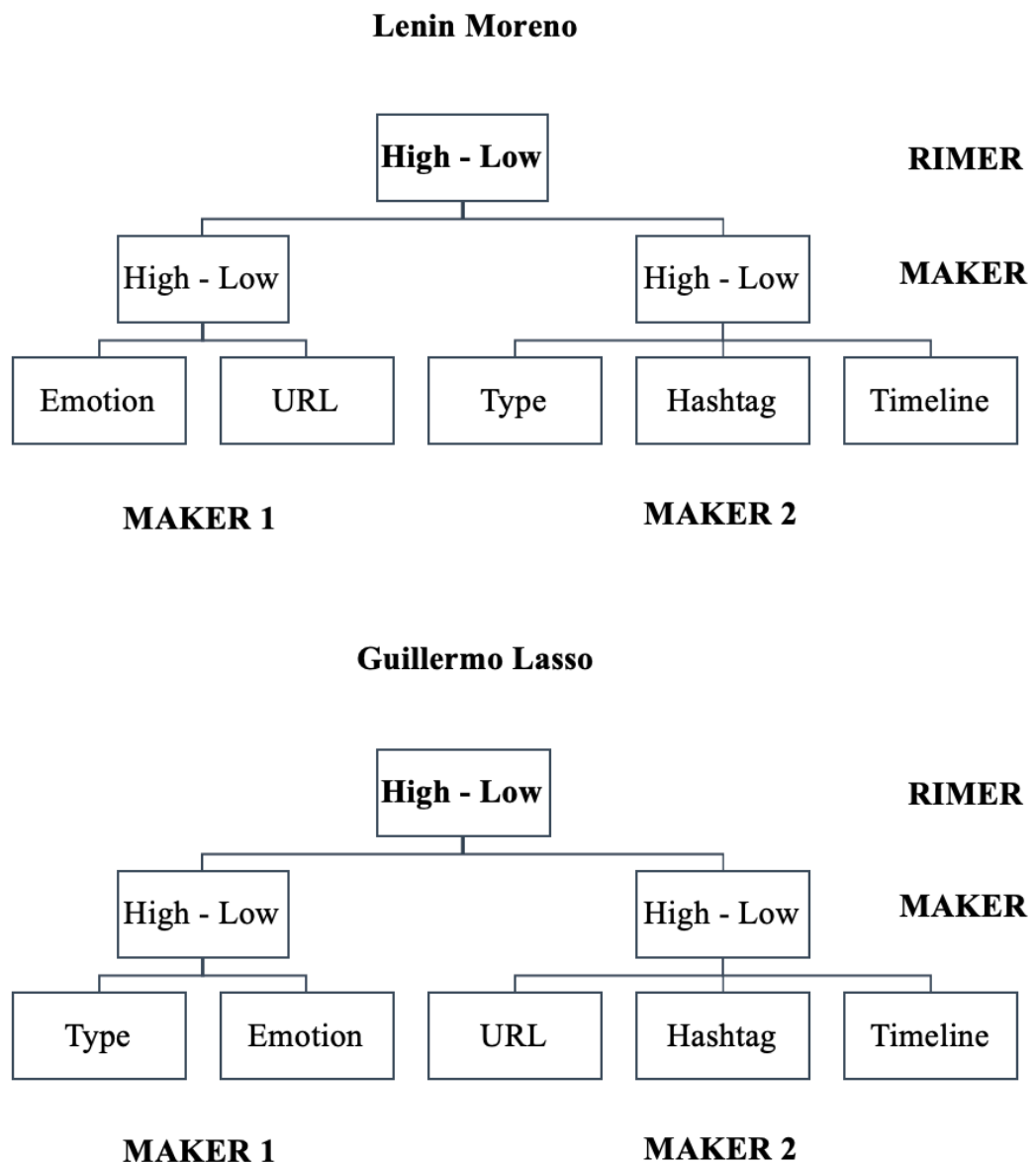


Figure 6.6. Hierarchical structures of the models for both candidates.

In the RIMER model, an initial belief rule base (BRB) is constructed, consisting of belief rules established on the basis of the types of outputs of the two

partial MAKER models after the training process as will be detailed in Subsection 6.5.3. For this reason, the parameters, which for RIMER are composed of the attribute weights of the two MAKER models and four belief rules, and the eight belief degrees of the four belief rules, need to be trained. The four belief rules come from the possible combination of the output, which can be High/High, High/Low, Low/High, and Low/Low. And the eight belief degrees are the results of all the possible consequents of a rule as shown in Table 6.6.

Table 6.6

Illustration of the possible belief rules and belief degrees used for this case study

Output	Four belief rules			
	High/High	High/Low	Low/High	Low/Low
High	Belief degree 1	Belief degree 3	Belief degree 5	Belief degree 7
Low	Belief degree 2	Belief degree 4	Belief degree 6	Belief degree 8

From the outputs of the two partial MAKER models, RIMER can be implemented as follows. The activation weight for each belief rule is calculated using Equation 6.7. Then, the degrees of belief β_{ik} are generated by implementing Equation 6.9. Following the training of the RIMER parameters, as will be explained in Subsection 6.5.3, the final RIMER results are presented in Table 6.7 for Lenin Moreno, and Table 6.8 for Guillermo Lasso. For example, from Table 6.8 about Guillermo Lasso, when MAKER 1 (i.e. type: announcement, and emotion: positive) is high, and MAKER 2 (i.e. URL: website, hashtag: not popular, and timeline: first) is low, the probability of a tweet being high impact is 0.6745, and of being low impact is 0.3255.

Figure 6.7 presents a generic model of the optimal learning process adapted from Yang et al. (2007). This model is used to represent the process to predict outcomes from input variables, which includes the implementation of the two partial MAKER models and RIMER methodology for each candidate.

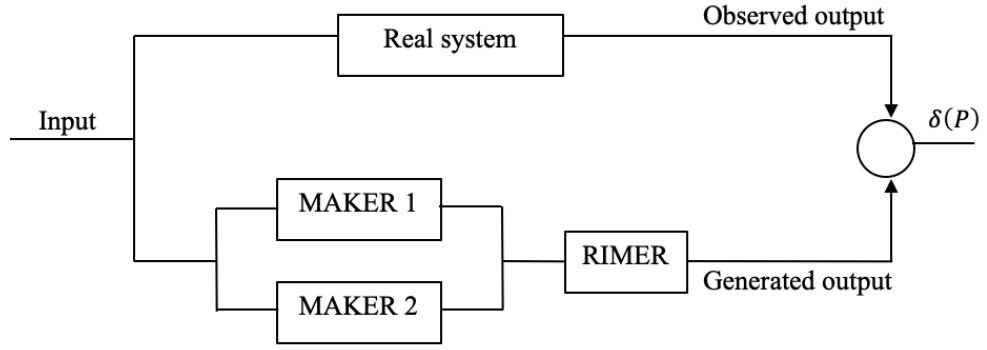


Figure 6.7. Illustration of the MAKER-RIMER generic training process.

6.5.3. Training the parameters of MAKER and RIMER

To begin with, weights are randomly assigned for both parameters of MAKER and those of the RIMER. The parameters for MAKER, for this study, are composed of the five input variables and their subcategories as presented in Figure 6.2. The parameters of the RIMER are the attributes, belief rules, and belief degrees. Then, weights need to be trained to improve the performance of the model (Xu et al., 2017; Yang et al., 2007), using the training dataset as introduced in Subsection 6.4.3. An optimisation model to minimise the misclassification error (MCE) is applied using the Equation 6.10, where P represents the vector of the parameters for MAKER and RIMER to be trained. The model above is optimised by means of Differential Evolution (Ardia, Mullen, Peterson, & Ulrich, 2016) as implemented in the R package "DEoptimR" (Conceicao & Maechler, 2016). The stopping criterion was based on the maximum number of iterations to be performed before the optimisation process is stopped, which was set to 5,000 iterations.

$$\delta(P) = \frac{1}{S} \sum_{s=1}^S \sum_{\theta \in \Theta} \left(p^{(s)}(\theta) - \hat{p}^{(s)}(\theta) \right)^2 \quad (6.10)$$

$$s. t. 0 \leq P \leq 1$$

where $\delta(P)$ refers to the objective function aiming to reduce the MCE, S is the total number of observations, $p^{(s)}(\theta)$ is the expected score of the output generated by using the MAKER and RIMER models for the s^{th} observation, and $\hat{p}^{(s)}(\theta)$ is the observed output of the s^{th} observation. The constraints of the training model for both MAKER and RIMER encompass normalisation of weights, so that they are between zero and one (Yang et al., 2007).

6.5.4. Validation with other machine learning approaches

Four statistical machine learning approaches were also applied to compare the results obtained using MAKER-RIMER. These approaches have demonstrated their effectiveness when dealing with Twitter data as detailed below: logistic regression (LR) (Culotta, 2010; Morgan, Lampe, & Shafiq, 2013), Naïve Bayes (NB) (Go, Huang, & Bhayani, 2009), decision tree (DT) (Castillo, Mendoza, & Poblete, 2011; Lee et al., 2011), and support vector machine (SVM) (Balabantaray, Mohammad, & Sharma, 2012).

The four machine learning approaches described above were also implemented in R, with the same training and testing datasets used to develop the MAKER-RIMER model. For this purpose, the following R packages were used: “nnet” (Venables & Ripley, 2002) for LR, “e1071” (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2017) for NB and SVM, and “tree” (Ripley, 2018) for DT. One difference when implementing these approaches compared to MAKER-RIMER is that for the former approaches the input variables were considered all at once in one single model and not hierarchically, as suggested in Figure 6.6.

To test the performance of the different classifiers, MCE was the metric used for comparison purposes. MCE, also known as error rate, refers to the total proportion of observations that are incorrectly classified across all the classes. This measure is used for evaluation since it works appropriately in predictive models with balanced outcome classes (Gu, Cai, Zhu, & Huang, 2008; Weiss, 2004), which for this case study comprises high and low impact. To calculate this metric, it is necessary to obtain the false positive (FP) and false negative (FN) results, which refer to the proportion of instances incorrectly classified for the considered classes as positive and negative respectively, while true positive (TP) and true negative (TN) refer to the proportion of instances correctly classified for the considered classes as positive and negative correspondingly, as shown in Equation 6.11.

$$MCE = \frac{FP + FN}{TP + FN + FP + TN} \quad (6.11)$$

6.6. Results and discussion

The results from MAKER-RIMER after training are presented in Table 6.7 for Lenin Moreno and Table 6.8 for Guillermo Lasso. These results were obtained using

the equations introduced in Subsections 6.2.1 and 6.2.2. The belief structures presented in both tables can help campaigners of political parties identify what features of tweets explain the achieving of high or low number of retweets. From an intuitive line of thinking, it can be assumed that when both partial MAKER models are high impact, which means that for both partial MAKER models the numbers of retweets are above the median of the total number of retweets, the result for RIMER should also be high impact. Similarly, when both partial MAKER models are low impact, meaning that the numbers of retweet of both partial MAKER models are below the median, the intuitive RIMER result is expected to be low impact as well. However, when the partial MAKER models are high/low or low/high, results from RIMER are uncertain. Hence, intuitive thinking might not be precise, so a robust model needs to be trained based on the actual data and weights.

The novelty of this approach, MAKER-RIMER, lies in that, to the best of the knowledge, it has not been applied before to experimental data. This approach allows to split a model into partial sub-models for analysis, and then combine them together. In this study, the MAKER-RIMER approach has dealt with limitation of quantity of data, which may lead to issues of uncertainty as stated in the introduction of this chapter. Moreover, the MAKER-RIMER approach, because of its interpretability, provides insights about the importance of features of tweets, namely type of tweet, emotion, URL, hashtag, and timeline, in the predicted tweet impact. Thus, the model can be used to reveal what features of a tweet are desirable, so it can achieve high impact. For example, using Guillermo Lasso's results, high impact in tweets is achieved either when both MAKER models are high (0.9651), or when the partial MAKER 1 model is high (0.6745). Similarly, low impact is obtained either when both MAKER results are low impact (0.8963), or when the partial MAKER 1 model is low (0.6797). Features of tweets leading to high and low impact are detailed at the end of this subsection.

Table 6.7

Rule base using RIMER with updated belief degrees considering MAKER 1 and MAKER 2 partial model results for Lenin Moreno

Antecedent	Consequent
(MAKER 1 is high \wedge MAKER 2 is high)	Impact of tweet is {(high, 0.9371), (low, 0.0629)}
(MAKER 1 is high \wedge MAKER 2 is low)	Impact of tweet is {(high, 0.3400), (low, 0.6600)}
(MAKER 1 is low \wedge MAKER 2 is high)	Impact of tweet is {(high, 0.3313), (low, 0.6687)}
(MAKER 1 is low \wedge MAKER 2 is low)	Impact of tweet is {(high, 0.2508), (low, 0.7492)}

Table 6.8

Rule base using RIMER with updated belief degrees considering MAKER 1 and MAKER 2 partial model results for Guillermo Lasso

Antecedent	Consequent
(MAKER 1 is high \wedge MAKER 2 is high)	Impact of tweet is {(high, 0.9651), (low, 0.0349)}
(MAKER 1 is high \wedge MAKER 2 is low)	Impact of tweet is {(high, 0.6745), (low, 0.3255)}
(MAKER 1 is low \wedge MAKER 2 is high)	Impact of tweet is {(high, 0.3203), (low, 0.6797)}
(MAKER 1 is low \wedge MAKER 2 is low)	Impact of tweet is {(high, 0.1037), (low, 0.8963)}

Therefore, the MAKER-RIMER approach proposed in this study provides several advantages. In terms of data availability, in this case study, the datasets do not contain all value combinations of input variables, thus we cannot create a single model that combines directly all five input variables. This limitation is overcome by developing two partial MAKER models and creating a hierarchical structure as shown in Figure 6.6. This is done by aggregating the input variables that are more closely correlated to form two partial MAKER models, as presented in Figure 6.6 and as explained in subsection 6.5.1. The MAKER-RIMER approach shows that, even when there are not possible combinations for all the input variables, the prediction is still evidence-based, and the reasoning is based on the knowledge of the data and not on intuition. Other data-driven modelling approaches might attempt to intuitively perform the prediction with all the variables together, even in the absence of all value combination of input variables to train the whole model. This may lead to models that are not fully interpretable and trusted because of the limitations of the data.

In terms of interpretability, the model presented in this research provides a robust procedure to map and represent the inputs and outputs (Kong et al., 2016; Yang et al., 2007). Unlike other machine learning approaches that do not provide a clear revelation of how the inference process unfolds, the MAKER-RIMER model shows transparently, in the partial MAKER models, the reasoning behind how variables are grouped together, and the weights assigned to the input variables and their subcategories. Also, it contemplates belief degrees, weights of antecedent attributes, and belief rules in RIMER (Yang et al., 2006). As explained in Subsection 6.5.3, the weights of parameters for MAKER and RIMER were first randomly assigned and later trained. In this sense, the initial assumption of equally weighted parameters is challenged because after the optimisation process, the weights of the parameters for the MAKER-RIMER model are trained to minimise the MCE when

predicting the output, which can be high or low impact. In summary, the MAKER-RIMER provides an evidence-based model, which shows an interpretable inference process that determines the outputs based on the available inputs from the data. Limitations of the model might arise especially when relationships between predictors and outcomes are not available, or prior knowledge is limited, so constructing the initial knowledge base represents a challenge (Kong et al., 2016; Yang et al., 2007). In addition, if the datasets are affected by noise, the generation of rules and the overall outcomes of MAKER-RIMER represents a challenge (Yang et al., 2006).

Lastly, the comparison of the performance of the different classification methods is shown in Table 6.9. These results suggest that the approach with the best performance in terms of minimum MCE is MAKER-RIMER for both candidates. MCE values for Guillermo Lasso were consistently smaller than those for Lenin Moreno because Guillermo Lasso posted almost twice as many tweets as Lenin Moreno.

Table 6.9

Comparison of performance of machine learning methods based on the MCE

Approaches	Lenin Moreno		Guillermo Lasso	
	MCE Train	MCE Test	MCE Train	MCE Test
MAKER-RIMER	0.4115	0.3385	0.2489	0.2373
LR	0.4250	0.3538	0.2574	0.2585
NB	0.4385	0.3923	0.2489	0.2500
DT	0.4635	0.4000	0.2532	0.2415
SVM	0.4269	0.4308	0.2595	0.2585

After obtaining results from the performance of the models, the next step is to identify which features of tweets affect the impact of tweets for each candidate. After applying the equations introduced in Section 6.2, the MAKER-RIMER results show that the two candidates share similar patterns when achieving the high impact of tweets, as presented in Figure 6.8 and Figure 6.9. Only combinations of variable values that give a probability of at least 0.75 of predicting high and at most 0.25 of predicting low are considered to predict high impact and, similarly, those value combinations that give a probability of at least 0.75 of predicting low and at most 0.25 of predicting high are considered to predict low impact. For example, the prominent high impact tweets include information about the contender, either with a positive or with a negative connotation. Tweets written during the last period have

higher impact. The presence of hashtags for both candidates is not associated with high impact. The difference between the candidates lies in URLs, since for Lenin Moreno high impact is linked with URLs about himself, while for Guillermo Lasso it is linked with URLs about other people or situations not directly related to his image. Concerning those tweets with low impact on retweets, it is observed that for Lenin Moreno, only one combination of the features of tweets generated low impact that comprises tweets containing announcements with sad connotations, having URLs about himself, without hashtags, and written in the last period of the campaign. However, for Guillermo Lasso, 25 possible combinations of features resulted in low impact. The most prevalent combinations included announcements with a neutral emotion, URLs containing information about himself or without URLs, with positive hashtags, and written in the first weeks of the campaign.

Impact	Type			Emotion			URL			Hashtag			Timeline	
	A	P	C	POS	NEG	SAD	H	O	N	P	NP	N	F	L
High														
Low														

Figure 6.8. Characteristics that make a tweet of high or low impact for Lenin Moreno• - See details of abbreviations of variables in Table A-6.1.

• Rows represent either high (green shading) or low (blue shading) impact of tweets in terms of number of retweets. Columns represent the variables, each one with its own parameters. So, each record denotes the possible combination of features of tweets to achieve the corresponding impact. For example, high impact for Lenin Moreno is achieved when the tweet is about the contender, with a negative emotion, having a URL about himself, without a hashtag, and written in the last period of the campaign.

Impact	Type			Emotion					URL				Hashtag			Timeline	
	A	P	C	POS	NEG	SAD	ANG	NEU	H	O	W	N	P	NP	N	F	L
High																	
Low																	

Figure 6.9. Characteristics that make a tweet of high or low impact for Guillermo Lasso[°] - See details of abbreviations of variables in Table A-6.1.

[°] Rows represent either high (green shading) or low (blue shading) impact of tweets in terms of number of retweets. Columns represent the variables, each one with its own parameters. So, each record denotes the possible combination of features of tweets to achieve the corresponding impact. For example, low impact for Guillermo Lasso happens when the tweet is an announcement, with a neutral emotion, having a URL about himself, without a hashtag, and written in the first period of the campaign.

Figure A-6.5 in the Appendix A-6 presents tweets labelled as high impact for both candidates. Lenin Moreno's tweet says "*The other candidate is not having a good time hugging poor people, because he is not used to it. He needs more solidarity*". Even though the tweet is not directly naming Guillermo Lasso, it is understood that the tweet is about him since he was the only contender during the second round of the election. The detected emotion is positive, the URL is about himself, there is no presence of hashtags, and it was written in the end of the campaign. Regarding Guillermo Lasso's tweet, the text says "*Very good @Lenin it was time to support the proposal of CHANGE. Because LASSO LASSO is CHANGE CHANGE*". The tweet is about the contender using the mention feature towards the other candidate, the emotion is positive, the URL shows an image not directly related to the candidate himself, it features no hashtags, and it was written in the last weeks of the campaign. Moreover, even though emoticons/emoji were not frequently used by the candidates throughout the campaign, this particular tweet used a "smiling face with sunglasses" emoji.

The results of this study are supported by previous research conducted in the political field. Concerning the diffusion of tweets based on type, those containing attacks on the contender tend to attract more attention from users (Darwish, Magdy, & Zanoluda, 2017; Lee & Xu, 2018). In addition, emotional content is more viral than non-emotional content (Berger & Milkman, 2012; González-Bailón et al., 2012). In this sense, a high diffusion of information is achieved when the content of tweets involves positive emotion (Stieglitz & Linh, 2012), but even higher if it conveys negative emotions (Choi, 2014; Lee & Xu, 2018). This study showed that high impact is associated with tweets involving either positive and negative emotions. Concerning URLs and hashtags, most studies have focused on the number thereof in tweets and their positive impact when retweeting, but surprisingly, for this study at least, the presence of hashtags is not prominently seen as vital when retweeting. Concerning the timeline, since the level of polarisation and Twitter content in the last period of the campaign increases, it is expected that tweets will gain more attention and diffusion (Cram, Llewellyn, Hill, & Magdy, 2017; Darwish et al., 2017).

6.7. Conclusion

This study has presented an ER based predictive model, MAKER-RIMER, to predict the impact of tweets in terms of the number of the retweets they achieved. The tweets posted by the two most voted candidates of the 2017 Ecuadorian Presidential election were used to develop the model. This model is based on likelihood data analysis and probabilistic inference via evidence combination. The proposed model provides a better interpretability of the reasoning process and results. It also presents and compares the performances of different machine learning approaches for prediction.

Findings show that the MAKER-RIMER model performed better than other machine learning approaches, namely logistic regression, Naïve Bayes, decision tree, and support vector machine, in predicting the impact of tweets, as it showed a smaller MCE. A smaller MCE is relevant since errors continue to be a barrier for machine learning approaches to be comprehensively adopted in prediction of human behaviour. The model presented in this study also allows the identification of features of a tweet that are predictors of its impact. The results have shown that for both candidates, high impact is obtained when their tweets include information about the contender, have either a positive or negative emotion, with URLs comprising information about the candidates themselves or about other people or situations not directly related to them, without the presence of hashtags, and written in the last period of the campaign.

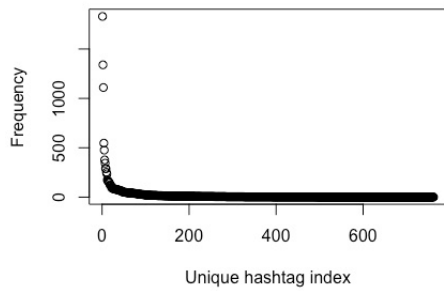
The generalisability of the results of this study is subject to certain limitations. This study only used tweets generated by the two most voted candidates to build the predictive model, with other users' tweets disregarded. In addition, the model is appropriate depending upon Twitter penetration among users and upon candidates' participation on Twitter. So, the model is appropriate when both parts generate and consume information on Twitter. In addition, the list of predictors in the model used in this study is not exhaustive. They depend on the fields and contexts of case studies, meaning that new variables can be added or adapted for consideration, which could be critical to outcomes, but can serve as a starting point for developing a retweeting model. For example, the proposed model does not include emoticons as input variables, because the candidates hardly used them throughout the campaign.

In terms of future research, the proposed MAKER-RIMER approach could be tested in different fields to analyse how the predictors in the model work, and how

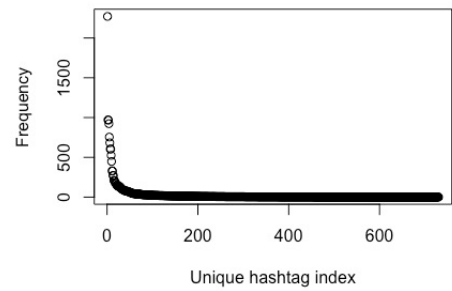
new variables can be adapted in different contexts. In addition, to enhance Twitter campaigns, future work could build retweeting predictive models which include tweets by other relevant users that generate high impact in terms of retweets. In this sense, candidates would be able to learn what works well for these users, so that they can adapt their own tweets accordingly. Another need for future research is the automation of variable classification, which was labelled manually in this study, by using machine learning, either via unsupervised (clustering) or supervised (classification) approaches. Finally, the model could be tested using unbalanced datasets for classifying high and low impact, for example assigning 30% of the data to high impact, while the remaining is assigned to low impact, to analyse the overall performance of the MAKER-RIMER model with imbalanced classes.

Appendix A-6.

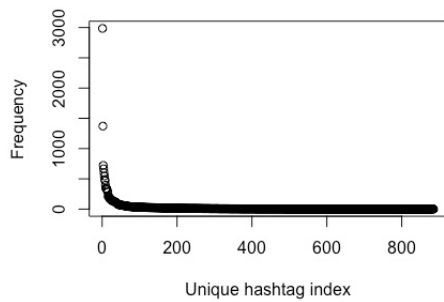
Week 1



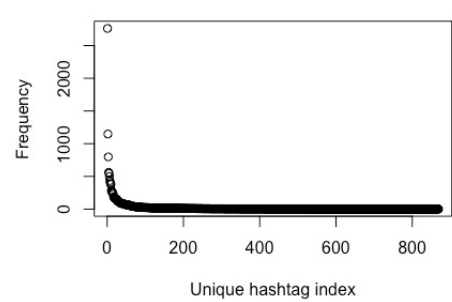
Week 2



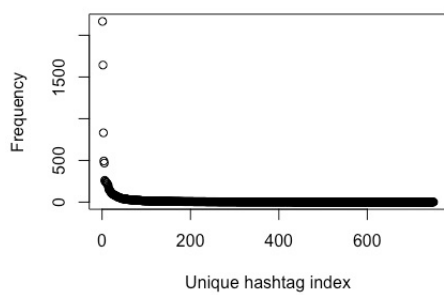
Week 3



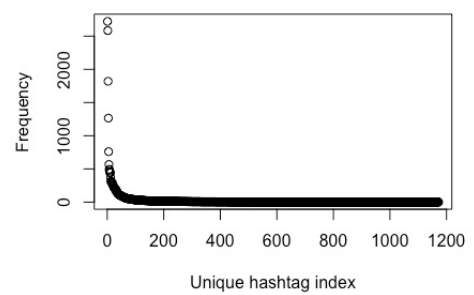
Week 4



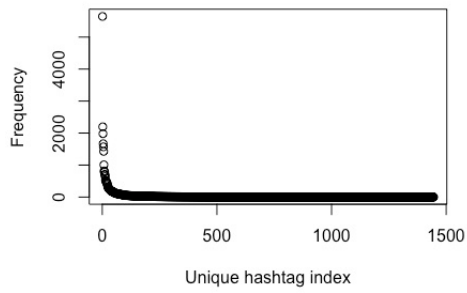
Week 5



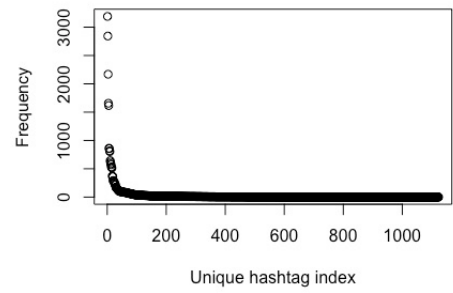
Week 6



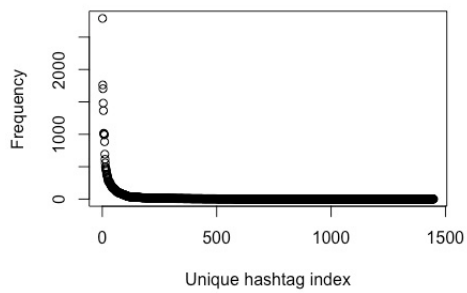
Week 7



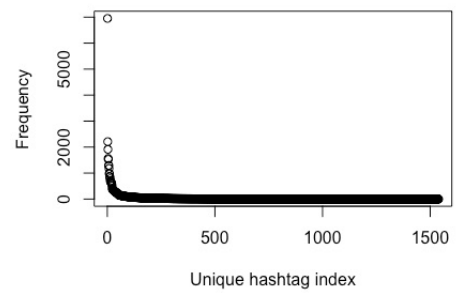
Week 8



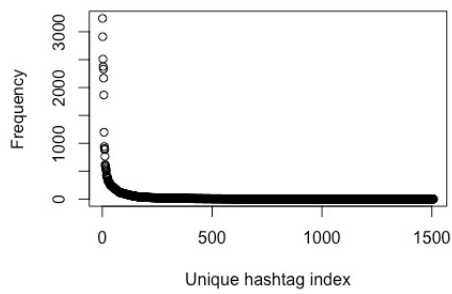
Week 9



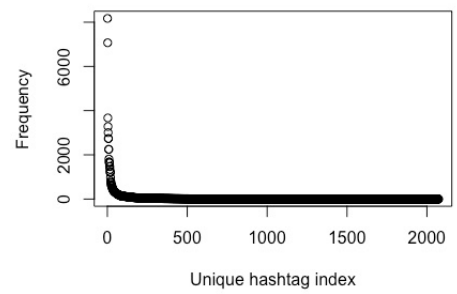
Week 10



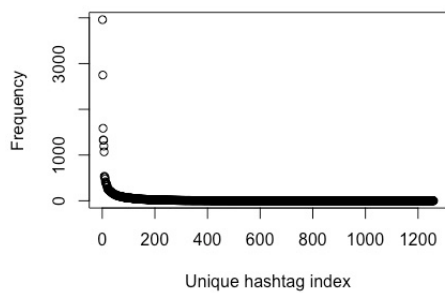
Week 11



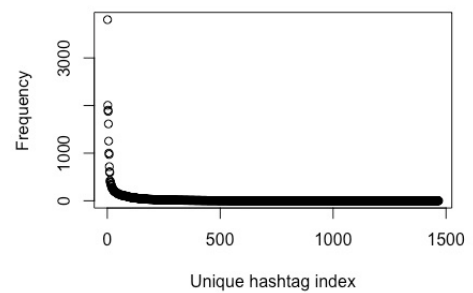
Week 12



Week 13



Week 14



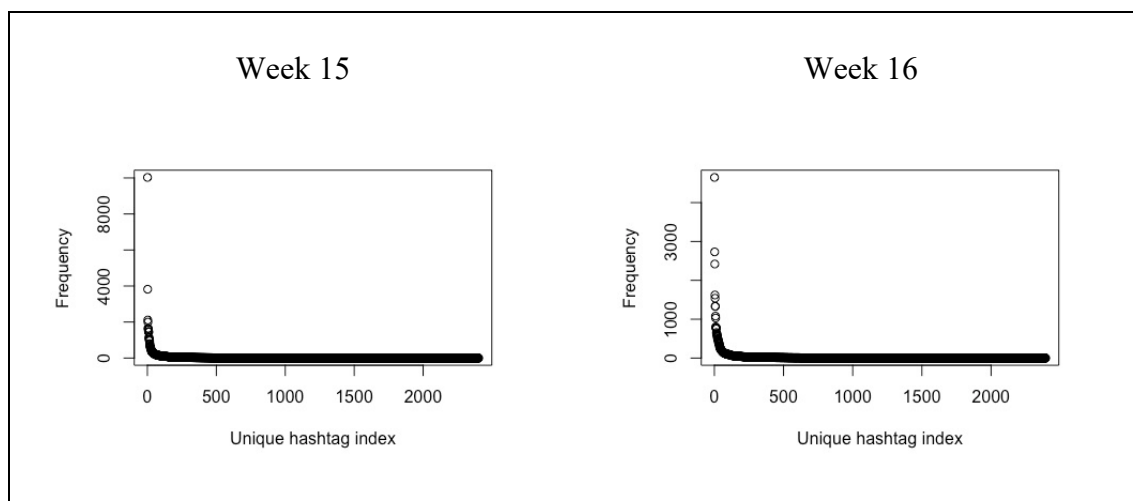
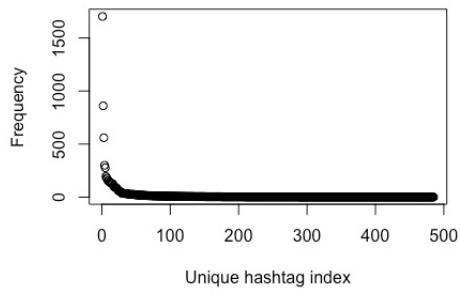
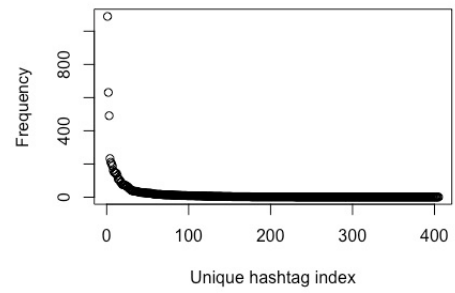


Figure A-6.1. Weekly breakdown of the frequency plot of hashtags found in tweets about Lenin Moreno.

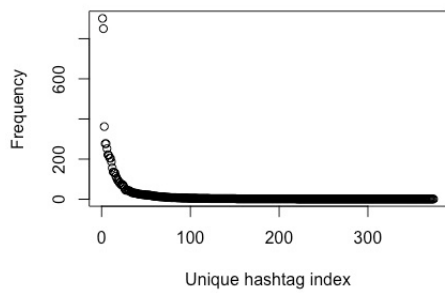
Week 1



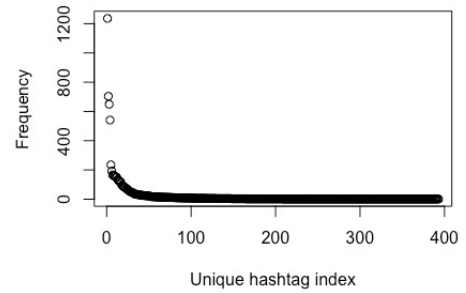
Week 2



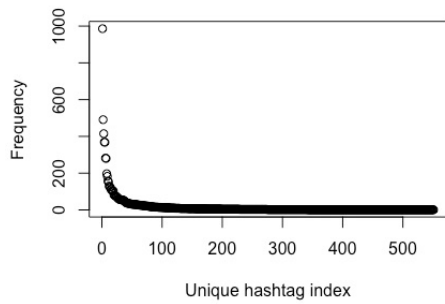
Week 3



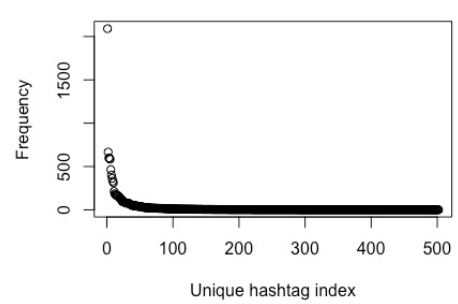
Week 4



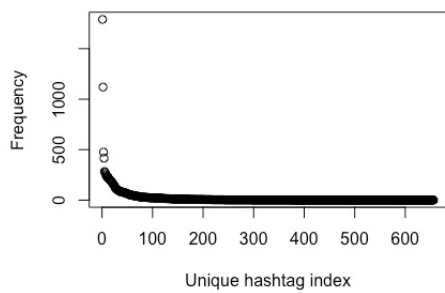
Week 5



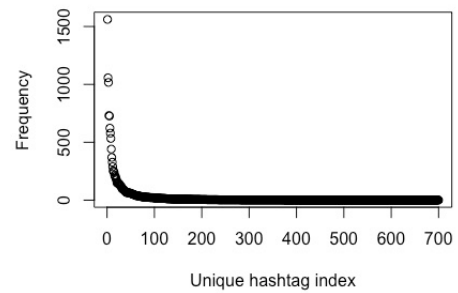
Week 6



Week 7



Week 8



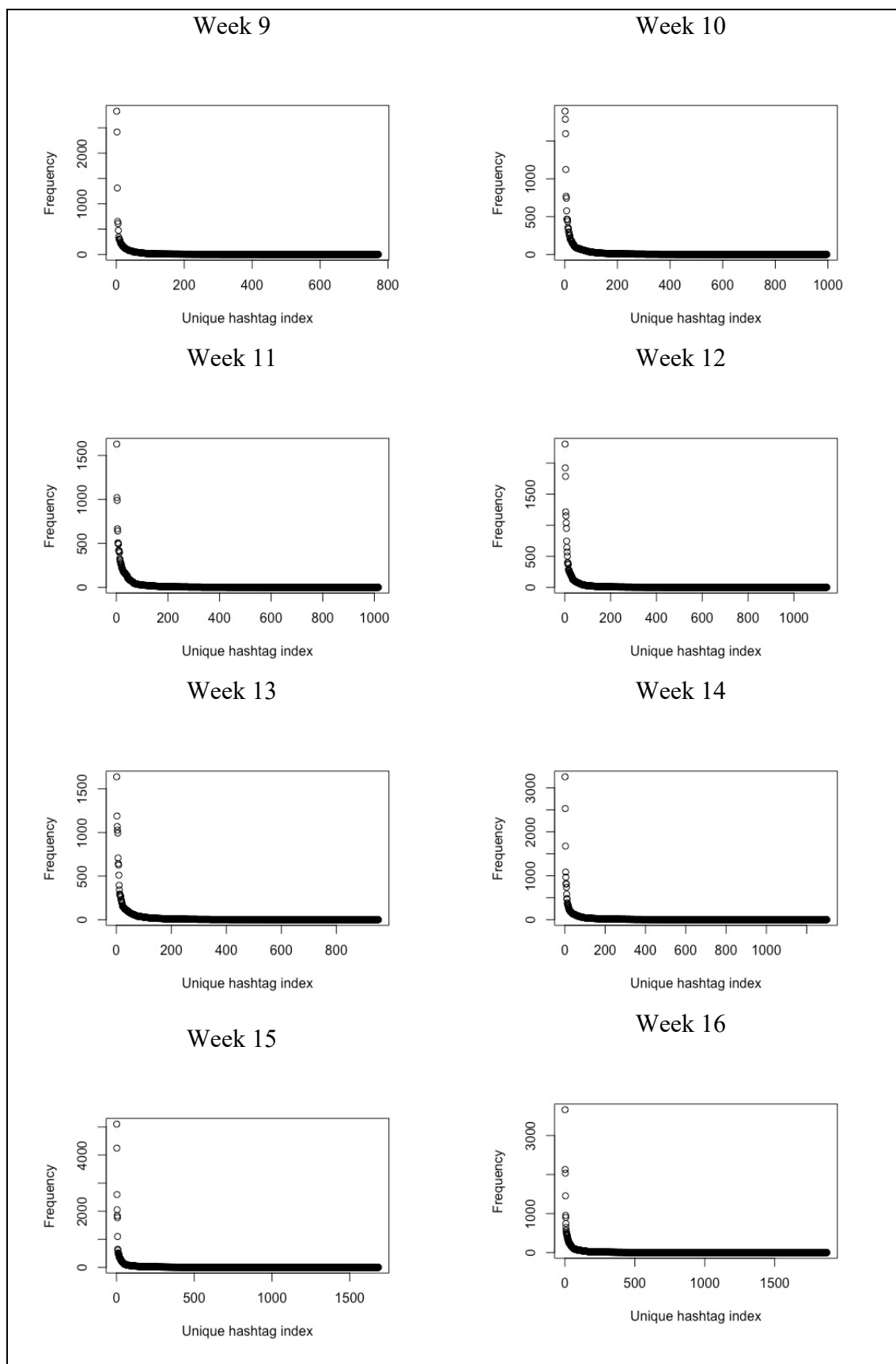


Figure A-6.2. Weekly frequency plot of hashtags found in tweets about Guillermo Lasso.

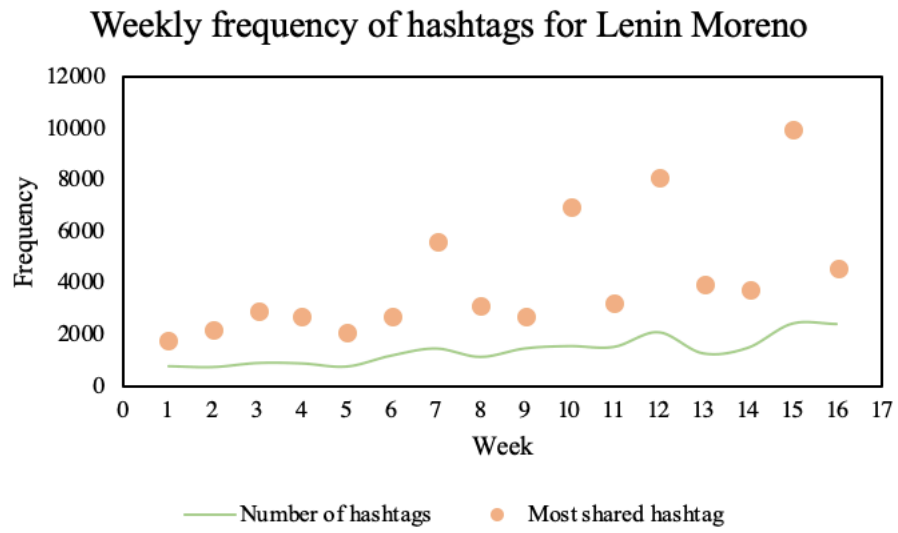


Figure A-6.3. Weekly frequency plot of hashtags for Lenin Moreno during the campaign.

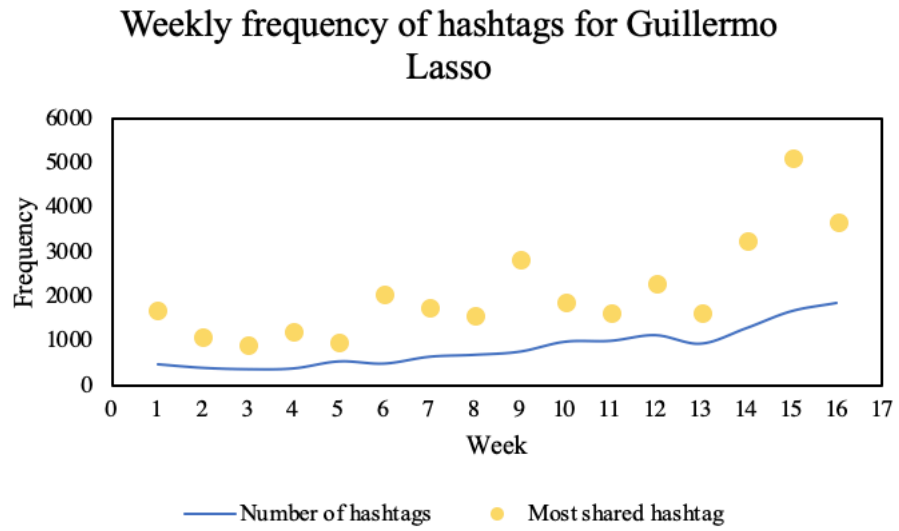


Figure A-6.4. Weekly frequency plot of hashtags for Guillermo Lasso during the campaign.



Figure A-6.5. Examples of tweets with high impact for Lenin Moreno and Guillermo Lasso.

Table A-6.1

Abbreviations of variables and their parameters used in this study

Abbreviations			Description
Type	A	Announcement	If the tweet contains information about facts or occurrences
	P	Proposal	If the tweet contains information about proposals
	C	Contender	If the tweet contains information about the other candidate
Emotion	POS	Positive	If the emotion contained in the tweet is positive
	NEG	Negative	If the emotion contained in the tweet is negative
	SAD	Sadness	If the emotion contained in the tweet is sad
	ANG	Anger	If the emotion contained in the tweet is angry
	NEU	Neutral	If no emotion can be detected in the tweet
URL	H	Himself	If the URL contain information about the candidate himself
	O	Other	If the URL contain information about other people/situations
	W	Website	If the URL redirect to an external website
	N	No URL	If no URL can be found in the tweet
Hashtag	P	Popular	If the hashtag is among the 20 most popular hashtags
	NP	Not popular	If the hashtag is not among the 20 most popular hashtags
	N	No hashtag	If no hashtag can be found in the tweet
Timeline	F	First	If the tweet was written between weeks 1 and 8
	L	Last	If the tweet was written between weeks 9 and 16

Results from Lenin Moreno

Table A-6.2

MAKER 1: Partial model involving emotion and URL for Lenin Moreno

Variables	MAKER 1 results after training weights	
	High	Low
POS-H	0.4963	0.5037
NEG-H	0.7400	0.2600
SAD-H	0.0376	0.9624
NEU-H	0.4081	0.5919
POS-O	0.5254	0.4746
NEG-O	0.9604	0.0396
SAD-O	0.5032	0.4968
NEU-O	0.5013	0.4987
POS-W	0.2112	0.7888
POS-N	0.5100	0.4900
NEG-N	0.6849	0.3151
SAD-N	0.6450	0.3550
NEU-N	0.6717	0.3283

Table A-6.3

MAKER 2: Partial model involving type, hashtag, and timeline for Lenin Moreno

Variables	MAKER 2 after training weights	
	High	Low
A-P-F	0.4061	0.5939
P-P-F	0.6349	0.3651
P-NP-F	0.9711	0.0289
A-N-F	0.4616	0.5384
P-N-F	0.6608	0.3392
A-P-L	0.6456	0.3544
P-P-L	0.7625	0.2375
A-NP-L	0.6803	0.3197
P-NP-L	0.4475	0.5525
A-N-L	0.4970	0.5030
P-N-L	0.4073	0.5927
C-N-L	1.0000	0.0000

Results from Guillermo Lasso

Table A-6.4

MAKER 2: Partial model involving URL, hashtag, and timeline for Guillermo Lasso

Variables	MAKER 2 results after training weights	
	High	Low
H-P-F	0.1295	0.8705
O-P-F	0.0900	0.9100
W-P-F	0.0107	0.9893
N-P-F	0.0712	0.9288
H-NP-F	0.0187	0.9813
O-NP-F	0.3496	0.6504
W-NP-F	0.1059	0.8941
N-NP-F	0.0485	0.9515
H-N-F	0.0756	0.9244
O-N-F	0.3149	0.6851
W-N-F	0.0153	0.9847
N-N-F	0.1348	0.8652
H-P-L	0.5561	0.4439
O-P-L	0.7942	0.2058
W-P-L	0.8208	0.1792
N-P-L	0.7739	0.2261
H-NP-L	0.4640	0.5360
W-NP-L	0.3495	0.6505
N-NP-L	0.8081	0.1919
H-N-L	0.6749	0.3251
O-N-L	0.8798	0.1202
W-N-L	0.4283	0.5717
N-N-L	0.7124	0.2876

References

- Allouch, N. (2018). Sentiment and emotional analysis: The absolute difference. Retrieved on October 15th, 2018 from <http://blog.emojics.com/emotional-analysis-vs-sentiment-analysis/>.
- André, P., Bernstein, M., & Luther, K. (2012). Who gives a tweet? Evaluating microblog content value. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (pp. 471-474). ACM, Seattle, Washington, US.
- Araujo, T., Neijens, P., & Vliegenthart, R. (2015). What motivates consumers to re-tweet brand content? The impact of information, emotion, and traceability on pass-along behavior. *Journal of Advertising Research*, 55(3), 284-295.
- Ardia, D., Mullen, K. M., Peterson, B. G., & Ulrich, J. (2016). "DEoptim": Differential Evolution in R. Version 2.2-4.
- Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone is an influencer: Quantifying influence on Twitter. In *Proceedings of the International Conference on Web Search and Data Mining* (pp. 65-74). ACM, Hong Kong.
- Balabantaray, R. C., Mohammad, M., & Sharma, N. (2012). Multi-class Twitter emotion classification: A new approach. *International Journal of Applied Information Systems*, 4(1), 48-53.
- Beck, M. B. (1987). Water quality modelling: A review of the analysis of uncertainty. *Journal of Water Resources Research*, 23(8), 1393-1442.
- Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, 49(2), 192-205.
- Bigsby, K. G., Ohlmann, J. W., & Zhao, K. (2019). The turf is always greener: Predicting decommitments in college football recruiting using Twitter data. *Journal of Decision Support Systems*, 116, 1-12.
- Bode, L., & Vraga, E. K. (2018). Studying politics across media. *Journal of Political Communication*, 35(1), 1-7.
- Boyd, D., Golder, S., & Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In *2010 International Conference on System Sciences* (pp. 1-10). IEEE, Honolulu, Hawaii, US.
- Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web* (pp. 675-684). ACM, Hyderabad, India.
- Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring user influence in Twitter: The million follower fallacy. In *Proceedings of the Fourth International AAAI Conference on Web and Social Media* (pp. 10-18). AAAI Press, Menlo Park, California, US.
- Chen, Y. W., Yang, J. B., Xu, D. L., Zhou, Z. J., & Tang, D. W. (2011). Inference analysis and adaptive training for belief rule-based systems. *Journal of Expert Systems with Applications*, 38(10), 12845-12860.
- Cheong, Y. P., & Gupta, R. (2005). Experimental design and analysis methods for assessing volumetric uncertainties. *SPE Journal*, 10(03), 324-335.
- Chiu, C.-Y., & Russell, A. D. (2011). Design of a construction management data visualization environment: A top-down approach. *Journal of Automation in Construction*, 20(4), 399-417.
- Choi, S. (2014). Flow, diversity, form, and influence of political talk in social-media-based public forums. *Journal of Human Communication Research*, 40(2), 209-237.

- Cogburn, D. L., & Espinoza-Vasquez, F. K. (2011). From networked nominee to networked nation: Examining the impact of Web 2.0 and social media on political participation and civic engagement in the 2008 Obama campaign. *Journal of Political Marketing*, 10(1-2), 189-213.
- Conceicao, E., & Maechler, M. (2016). Differential Evolution Optimization in Pure R [RStudio package].
- Cram, L., Llewellyn, C., Hill, R., & Magdy, W. (2017). UK general election 2017: A Twitter analysis. *arXiv preprint arXiv:1706.02271*.
- Culotta, A. (2010). Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the First Workshop on Social Media Analytics* (pp. 115-122). ACM, Washington, US.
- Cunha, E., Magno, G., Comarela, G., Almeida, V., Gonçalves, M. A., & Benevenuto, F. (2011). Analyzing the dynamic evolution of hashtags on Twitter: A language-based approach. In *Proceedings of the Workshop on Languages in Social Media* (pp. 58-65). Association for Computational Linguistics, Pennsylvania, US.
- Dabeer, O., Mehendale, P., Karnik, A., & Saroop, A. (2011). Timing tweets to increase effectiveness of information campaigns. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (pp. 105-112). AAAI Press, Barcelona, Spain.
- Darwish, K., Magdy, W., & Zanoluda, T. (2017). Trump vs. Hillary: What went viral during the 2016 US presidential election. *arXiv preprint arXiv :1707.03375*.
- De Veirman, M., Cauberghe, V., & Hudders, L. (2017). Marketing through Instagram influencers: The Impact of number of followers and product divergence on brand attitude. *International Journal of Advertising*, 36(5), 798-828.
- Dempster, A. (1967). Upper and Lower Probabilities Induced by a Multivalued Mapping. *The Annals of Mathematical Statistics*, 38(2), 325-339.
- Dubois, E., & Gaffney, D. (2014). The multiple facets of influence: Identifying political influentials and opinion leaders on Twitter. *Journal of American Behavioral Scientist*, 58(10), 1260-1277.
- Ekman, P. (1993). Facial expression and emotion. *Journal of American Psychologist*, 48(4), 384-392.
- Enli, G. (2017). Twitter as arena for the authentic outsider: Exploring the social media campaigns of Trump and Clinton in the 2016 US Presidential election. *European Journal of Communication*, 32(1), 50-61.
- Entman, R. M. (1992). Blacks in the news: Television, modern racism and cultural change. *Journalism Quarterly*, 69(2), 341-361.
- Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research*, 13(4), 1-20.
- Fan, W.-J., Yang, S.-L., Perros, H., & Pei, J. (2015). A multi-dimensional trust-aware cloud service selection mechanism based on evidential reasoning approach. *International Journal of Automation and Computing*, 12(2), 208-219.
- Feng, W., & Wang, J. (2013). Retweet or not? Personalized tweet re-ranking. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining* (pp. 577-586). ACM, Rome, Italy.
- Fogarty, B. J., & Wolak, J. (2009). The effects of media interpretation for citizen evaluations of politicians' messages. *Journal of American Politics Research*, 37(1), 129-154.
- Go, A., Huang, L., & Bhayani, R. (2009). Twitter sentiment analysis. *Journal of Entropy*, 17, 252-258.

- González-Bailón, S., Banchs, R. E., & Kaltenbrunner, A. (2012). Emotions, public opinion, and US presidential approval rates: A 5-year analysis of online political discussions. *Journal of Human Communication Research*, 38(2), 121-143.
- Gu, Q., Cai, Z., Zhu, L., & Huang, B. (2008). Data mining on imbalanced data sets. In *2008 International Conference on Advanced Computer Theory and Engineering* (pp. 1020-1024). IEEE, Phuket, Thailand.
- Hammami, M., Chahir, Y., & Chen, L. (2003). WebGuard: Web based adult content detection and filtering system. In *2003 International Conference on Web Intelligence* (pp. 574-578). IEEE, Halifax, Canada.
- Hemphill, L., Culotta, A., & Heston, M. (2013). Framing in social media: How the US Congress uses Twitter hashtags to frame political issues. *Available at SSRN 2317335*. <http://dx.doi.org/10.2139/ssrn.2317335>.
- Hochreiter, R., & Waldhauser, C. (2013). A stochastic simulation of the decision to retweet. In *2013 International Conference on Algorithmic Decision Theory* (pp. 221-229). Springer, Bruxelles, Belgium.
- Hong, L., Dan, O., & Davison, B. D. (2011). Predicting popular messages in Twitter. In *Proceedings of the 20th International Conference Companion on World Wide Web* (pp. 57-58). ACM, Hyderabad, India.
- Huang, D., Zhou, J., Mu, D., & Yang, F. (2014). Retweet behavior prediction in Twitter. In *2014 Seventh International Symposium on Computational Intelligence and Design* (pp. 30-33). IEEE, Hangzhou, China.
- Jaidka, K., Ahmed, S., Skoric, M., & Hilbert, M. (2018). Predicting elections from social media: A three-country, three-method comparative study. *Asian Journal of Communication*, 29(3), 1-21.
- Jin, S.-A. A., & Phua, J. (2014). Following celebrities' tweets about brands: The impact of Twitter-based electronic word-of-mouth on consumers' source credibility perception, buying intention, and social identification with celebrities. *Journal of Advertising*, 43(2), 181-195.
- Kaur, J., & Saini, J. R. (2014). Emotion detection and sentiment analysis in text corpus: A differential study with informal and formal writing styles. *International Journal of Computer Applications*, 101(9), 1-9.
- Kavuluru, R., & Sabbir, A. K. M. (2016). Toward automated e-cigarette surveillance: Spotting e-cigarette proponents on Twitter. *Journal of Biomedical Informatics*, 61, 19-26.
- Keib, K., Himelboim, I., & Han, J. Y. (2018). Important tweets matter: Predicting retweets in the #BlackLivesMatter talk on Twitter. *Journal of Computers in Human Behavior*, 85, 106-115.
- Kim, E., Hou, J., Han, J. Y., & Himelboim, I. (2016). Predicting retweeting behavior on breast cancer social networks: Network and content characteristics. *Journal of Health Communication*, 21(4), 479-486.
- Kim, Y. K., Lee, D., Lee, J., Lee, J. H., & Straub, D. W. (2018). Influential users in social network services: The contingent value of connecting user status and brokerage. *Journal of ACM SIGMIS Database: The Database for Advances in Information Systems*, 49(1), 13-32.
- Kim, E., Sung, Y., & Kang, H. (2014). Brand followers' retweeting behavior on Twitter: How brand relationships influence brand electronic word-of-mouth. *Journal of Computers in Human Behavior*, 37, 18-25.
- Kong, G., Xu, D.-L., Yang, J.-B., & Ma, X. (2015). Combined medical quality assessment using the evidential reasoning approach. *Journal of Expert Systems with Applications*, 42(13), 5522-5530.

- Kong, G., Xu, D.-L., Yang, J.-B., Yin, X., Wang, T., Jiang, B., & Hu, Y. (2016). Belief rule-based inference for predicting trauma outcome. *Journal of Knowledge-Based Systems*, 95, 35-44.
- Kwakkel, J. H., Walker, W. E., & Haasnoot, M. (2016). Coping with the wickedness of public policy problems: Approaches for decision making under deep uncertainty. *Journal of Water Resources Planning and Management*, 142(3), 1-5.
- Laugel, T., Lesot, M. J., Marsala, C., Renard, X., & Detyniecki, M. (2018). Comparison-based inverse classification for interpretability in machine learning. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (pp. 100-111). Springer, Cham, Cádiz, Spain.
- Lee, J., & Lim, Y.-S. (2016). Gendered campaign tweets: The cases of Hillary Clinton and Donald Trump. *Journal of Public Relations Review*, 42(5), 849-855.
- Lee, J., & Xu, W. (2018). The more attacks, the more retweets: Trump's and Clinton's agenda setting on Twitter. *Journal of Public Relations Review*, 44(2), 201-213.
- Lee, K., Palsetia, D., Narayanan, R., Patwary, M. M. A., Agrawal, A., & Choudhary, A. (2011). Twitter trending topic classification. In *11th International Conference on Data Mining Workshops* (pp. 251-258). IEEE, Vancouver, British Columbia, Canada.
- Lee, M., Kim, H., & Kim, O. (2015). Why do people retweet a tweet? Altruistic, egoistic, and reciprocity motivations for retweeting. *Journal of Psychologia*, 58(4), 189-201.
- Lim, K. Y. (2018). An exploration of the use of Facebook by legislators in Taiwan. *Journal of Issues & Studies*, 54(03), 1840005.
- Lin, W.-H., Green, T. H., Kaplow, R., Fu, G., & Mann, G. S. (2012). Predictive model importation: Google Patents.
- Lo, S. L., Chiong, R., & Cornforth, D. (2016). Ranking of high-value social audiences on Twitter. *Journal of Decision Support Systems*, 85, 34-48.
- Luo, Z., Osborne, M., Tang, J., & Wang, T. (2013). Who will retweet me? Finding retweeters in Twitter. In *Proceedings of the 36th International ACM Conference on Research and Development in Information Retrieval* (pp. 869-872). ACM, Dublin, Ireland.
- Macskassy, S. A., & Michelson, M. (2011). Why do people retweet? Anti-homophily wins the day!. In *Proceedings of the Fifth International AAAI Conference on Web and Social Media* (pp. 209-216). AAAI Press, Barcelona, Spain.
- Mahadevan, S., & Sarkar, S. (2009). Uncertainty analysis methods. *US Department of Energy, Washington, DC, US*.
- Maharana, B., Nayak, K., & Sahu, N. K. (2006). Scholarly use of web resources in LIS research: A citation analysis. *Library Review*.
- Maity, S. K., Gajula, R., & Mukherjee, A. (2018). Why did they unfollow me? Early detection of follower loss on Twitter. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork* (pp. 127-131). ACM, Florida, US.
- McDonald, J. (2014). Small numbers in chi-square and G-tests. *Handbook of Biological Statistics*, 86-89.
- Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2017). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-8. <https://CRAN.R-project.org/package=e1071>.

- Morgan, J. S., Lampe, C., & Shafiq, M. Z. (2013). Is news sharing on Twitter ideologically biased?. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (pp. 887-896). ACM, San Antonio, Texas, US.
- Muggler, M., Eshwarappa, R., & Cankaya, E. C. (2017). Cybersecurity management through logging analytics. In *International Conference on Applied Human Factors and Ergonomics* (pp. 3-15). Springer, Cham, Los Angeles, US.
- Nagaonkar, A. R., & Kulkarni, U. L. (2016). Finding the malicious URLs using search engines. In *3rd International Conference on Computing for Sustainable Global Development* (pp. 3692-3694). IEEE, New Delhi, India.
- Nesi, P., Pantaleo, G., Paoli, I., & Zaza, I. (2018). Assessing the retweet proneness of tweets: Predictive models for retweeting. *Journal of Multimedia Tools and Applications*, 77(20), 26371–26396.
- Noriega, M. (2014). Why we retweet? *The Daily Dot*. Retrieved on August 8th, 2018 from <https://www.dailydot.com/debug/why-we-retweet/>.
- Ott, B. L. (2017). The age of Twitter: Donald J. Trump and the politics of debasement. *Journal of Critical Studies in Media Communication*, 34(1), 59-68.
- Parmelee, J., & Bichard, S. (2011). In their own words: Exploring the role and value of political Twitter use in followers' lives. In J. H. Parmelee & S. L. Bichard (Eds.), *Politics and the Twitter revolution: How tweets influence the relationship* (pp. 142-166): Lexington Books.
- Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. *The University of Texas at Austin*.
- Petrovic, S., Osborne, M., & Lavrenko, V. (2011). RT to win! Predicting message propagation in Twitter. In *Proceedings of the Fifth International AAAI Conference on Web and Social Media* (pp. 586-589). AAAI Press, Barcelona, Spain.
- Pezzoni, F., An, J., Passarella, A., Crowcroft, J., & Conti, M. (2013). Why do I retweet it? An information propagation model for microblogs. In *International Conference on Social Informatics* (pp. 360-369). Springer, Kyoto, Japan.
- R Core Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (Version 1.0.136). Retrieved from <http://www.r-project.org/>.
- Ramos-Serrano, M., Fernandez-Gomez, J. D., & Pineda, A. (2018). "Follow the closing of the campaign on streaming": The use of Twitter by Spanish political parties during the 2014 European elections. *Journal of New Media & Society*, 20(1), 122-140.
- Ripley, B. (2018). tree: Classification and regression trees. R package version 1.0-39. <https://CRAN.R-project.org/package=tree>.
- Rolnick, D., Veit, A., Belongie, S., & Shavit, N. (2017). Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*.
- Rudat, A., & Buder, J. (2015). Making retweeting social: The influence of content and context information on sharing news in Twitter. *Journal of Computers in Human Behavior*, 46, 75-84.
- Sabin, A., Xu, D. L., Chen, Y. W., & Savan, E. E. (2013). An evaluation of website upgrade options: A case study comparison of ANFIS and RIMER. In *2013 International Conference on Systems, Man, and Cybernetics* (pp. 1385-1390). IEEE, Washington DC, US.
- Savage, S., Monroy-Hernandez, A., & Höllerer, T. (2016). Botivist: Calling volunteers to action using online bots. In *Proceedings of the 19th ACM*

- Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 813-822). ACM, San Francisco, California, US.
- Shafer, G. (1976). *A mathematical theory of evidence* (Vol. 42): Princeton University Press.
- Smarandache, F., Dezert, J., & Tacnet, J.-M. (2010). Fusion of sources of evidence with different importances and reliabilities. In *Workshop on the Theory of Belief Functions* (pp. 1-9). Edinburgh, UK.
- Soulier, L., Tamine, L., & Nguyen, G. H. (2016). Answering Twitter questions: A model for recommending answerers through social collaboration. In *Proceedings of the 25th International Conference on Information and Knowledge Management* (pp. 267-276). ACM, New York, US.
- Srivastava, R. P. (2011). An introduction to evidential reasoning for decision making under uncertainty: Bayesian and belief function perspectives. *International Journal of Accounting Information Systems*, 12(2), 126-135.
- Stieglitz, S., & Linh, D.-X. (2012). Political communication and influence through microblogging. An empirical analysis of sentiment in Twitter messages and retweet behavior. In *45th Hawaii International Conference on System Science* (pp. 3500-3509). IEEE, Maui, Hawaii, US.
- Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010). Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In *2010 IEEE International Conference on Social Computing* (pp. 177-184). IEEE, Minneapolis, Minnesota, US.
- Tamaddoni, A., Stakhovych, S., & Ewing, M. (2017). The impact of personalised incentives on the profitability of customer retention campaigns. *Journal of Marketing Management*, 33(5-6), 327-347.
- Tan, C., Lee, L., & Pang, B. (2014). The effect of wording on message propagation: Topic-and author-controlled natural experiments on Twitter. *arXiv preprint arXiv:1405.1438*.
- Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. *Journal of PLOS ONE*, 8(5), e64841.
- Trilling, D., Tolochko, P., & Burscher, B. (2017). From newsworthiness to shareworthiness: How to predict news sharing based on article characteristics? *Journalism & Mass Communication Quarterly*, 94(1), 38-60.
- Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. *Journal of The International Linguistic Association*, 10(1), 178-185.
- Varghese, A., Cawley, M., & Hong, T. (2018). Supervised clustering for automated document classification and prioritization: A case study using toxicological abstracts. *Journal of Environment Systems and Decisions*, 38(3), 398-414.
- Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S*. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.
- Wang, M., Zuo, W., & Wang, Y. (2015). A multidimensional nonnegative matrix factorization model for retweeting behavior prediction. *Journal of Mathematical Problems in Engineering*, 2015, 1-11.
- Wang, Y., Luo, J., Niemi, R., Li, Y., & Hu, T. (2016). Catching fire via "likes": Inferring topic preferences of Trump followers on Twitter. In *Proceedings of the Tenth International AAI Conference on Web and Social Media* (pp. 719-722). AAAI Press, Cologne, Germany.

- Wang, Y.-M., Yang, J.-B., & Xu, D.-L. (2006). Environmental impact assessment using the evidential reasoning approach. *European Journal of Operational Research*, 174(3), 1885-1913.
- Weingart, P. (2007). Communicating science in democratic media societies. *Journal of Communicating Global Change Science to Society: An Assessment and Case Studies*, 68, 55.
- Weiss, G. M. (2004). Mining with rarity: A unifying framework. *Journal of ACM SIGKDD Explorations*, 6(1), 7-19.
- Wells, C., Shah, D. V., Pevehouse, J. C., Yang, J., Pelled, A., Boehm, F., . . . Schmidt, J. L. (2016). How Trump drove coverage to the nomination: Hybrid media campaigning. *Journal of Political Communication*, 33(4), 669-676.
- Wong, W. S., Amer, M., Maul, T., Liao, I. Y., & Ahmed, A. (2020). Conditional generative adversarial networks for data augmentation in breast cancer classification. In *International Conference on Soft Computing and Data Mining* (pp. 392-402). Springer, Cham, Putrajaya, Malaysia.
- Xu, D.-L., & Yang, J.-B. (2006). Intelligent decision system and its application in business innovation self assessment. *Journal of Decision Support Systems*, 42(2), 664-673.
- Xu, X., Zhao, Z., Xu, X., Yang, J., Chang, L., Yan, X., & Wang, G. (2020). Machine learning-based wear fault diagnosis for marine diesel engine by fusing multiple data-driven models. *Journal of Knowledge-Based Systems*, 190, 105324.
- Xu, X., Zheng, J., Yang, J.-b., Xu, D.-l., & Chen, Y.-w. (2017). Data classification using evidence reasoning rule. *Journal of Knowledge-Based Systems*, 116, 144-151.
- Xu, Z., & Yang, Q. (2012). Analyzing user retweet behavior on Twitter. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 46-50). IEEE, Istanbul, Turkey.
- Yang, J.-B. (2001). Rule and utility based evidential reasoning approach for multiattribute decision analysis under uncertainties. *European Journal of Operational Research*, 131(1), 31-61.
- Yang, J.-B., Liu, J., Wang, J., Sii, H.-S., & Wang, H.-W. (2006). Belief rule-base inference methodology using the evidential reasoning approach - RIMER. *IEEE Transactions on systems, Man, and Cybernetics-part A: Systems and Humans*, 36(2), 266-285.
- Yang, J.-B., Liu, J., Xu, D.-L., Wang, J., & Wang, H. (2007). Optimization models for training Belief-Rule-Based systems. *IEEE Transactions on systems, Man, and Cybernetics-part A: Systems and Humans*, 37(4), 569-585.
- Yang, J.-B., & Xu, D.-L. (2013). Evidential reasoning rule for evidence combination. *Journal of Artificial Intelligence*, 205, 1-29.
- Yang, J.-B., & Xu, D.-L. (2014). A study on generalising Bayesian inference to evidential reasoning. In *2014 International Conference on Belief Functions: Theory and Applications* (pp. 180-189). Springer, Oxford, UK.
- Yang, J.-B., & Xu, D.-L. (2017). Inferential modelling and decision making with data. In *2017 International Conference on Automation and Computing* (pp. 1-6). IEEE, Huddersfield, UK.
- Yang, Z., Guo, J., Cai, K., Tang, J., Li, J., Zhang, L., & Su, Z. (2010). Understanding retweeting behaviors in social networks. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (pp. 1633-1636). ACM, Toronto, Ontario, Canada.
- Yates, D., Moore, D., McCabe, G. (1999). *The Practice of Statistics (1st Ed.)*.

- Zadeh, A. H., & Sharda, R. (2014). Modeling brand post popularity dynamics in online social networks. *Journal of Decision Support Systems*, 65, 59-68.
- Zandona, F. P., Rault, S. S. J. M., & Ripsher, L. B. (2013). Temporal topic extraction. U.S. Patent Application No. 13/530,495.
- Zhang, J., Jiang, J., Chen, Y., & Yang, K. (2015). A belief rule-based safety evaluation approach for complex systems. *Journal of Mathematical Problems in Engineering*, 2015.
- Zhang, D., Yan, X., Zhang, J., Yang, Z., & Wang, J. (2016). Use of fuzzy rule-based evidential reasoning approach in the navigational risk assessment of inland waterway transportation systems. *Journal of Safety Science*, 82, 352-360.
- Zhong, C., Batty, M., Manley, E., Wang, J., Wang, Z., Chen, F., & Schmitt, G. (2016). Variability in regularity: Mining temporal mobility patterns in London, Singapore and Beijing using smart-card data. *Journal of PLOS ONE*, 11(2), e0149222.
- Zhu, H., Yang, J.-B., Xu, D.-L., & Xu, C. (2016). Application of Evidential Reasoning rules to identification of asthma control steps in children. In *22nd International Conference on Automation and Computing* (pp. 444-449). IEEE, Colchester, UK.

Chapter 7

Conclusion and future research path

This thesis set out to develop and apply new methods for the analysis of Twitter data that are suitable for political context. This work is timely and relevant in light of the surge of interest in the way Twitter supports the understanding of user behaviours. This concluding section is organised in four parts. It begins by summarising the main arguments and findings developed in this thesis. The second part highlights the key contributions and implications of this study. The third part reflects on the strengths and limitations. Lastly, the fourth part suggests some directions for future research.

7.1. Towards a novel approach to Twitter analysis

This section sums up the thesis and highlights its main arguments. This thesis is organised into three academic papers. The first paper, presented in Chapter 4, has reviewed the literature on Twitter use in the political context to analyse and understand user behaviour. The emergence of Twitter has brought new data sources worthy of consideration because it furthers understanding of how users make choices. The other two papers have aimed to understand what Twitter can tell us about users involved in political conversations during electoral processes. Twitter data can support campaigners of political parties and candidates in two ways. One way is to have an indication of vote choice during electoral processes, and thereby an anticipation of vote share. Chapter 5 presented two new approaches that enhance the ability of sentiment analysis tools to detect sentiment on Twitter and identify influential users which could trigger the sentiment. These approaches provide new ways to estimate vote share and identify sources of support and opposition, for which they constitute progress in making Twitter data relevant in the political context. The other way it to identify what engages users on Twitter to retweet candidates' tweets during an electoral campaign. Chapter 6 presented a novel predictive model that can help candidates recognise features on their tweets leading to high or low impact in terms of the number of retweets achieved.

The literature on the use of Twitter for analysis of user behaviour still has a long way to go before reaching maturity. As shown in Chapter 4, some of the key issues that remain unresolved include, first of all, a clear set of guidelines that can be

used to deal with the sampling process in social media. On this matter, the literature subtly indicates that deciding on the keywords/users to extract tweets, the number of tweets needed for analysis, as well as the timeframe involved, has been done essentially intuitively rather than being empirically driven. Furthermore, the sampling aspect is perhaps the main criticism and weakness attributed to Twitter data in the literature. Secondly, the way of using features in tweets, such as hashtags, emoticons, and URLs. In most of the studies reviewed, these features have been discarded from the analysis or simplistically coded, which might limit the understanding of users' needs and interests. While this study has offered insights, there is still work to be done to better understand the information and sentiment these features transmit. Thirdly, there is a lack of mechanisms that can be used to assess the impact of fake users and misleading information on sentiment. Regarding this, much of the literature has focused on methods to identify the veracity of an account or news, while the pertinent impact remains under-explored. Lastly, new methods need to be developed to identify influential users and measure their impact on the sentiment or opinions towards a given entity on Twitter.

The other two papers presented in Chapter 5 and Chapter 6 concern some of the gaps identified above. The second paper proposed two approaches to enhance the functionality of Twitter data. First, an approach for pre-processing tweets is intended for sentiment analysis, including coding hashtags, emoticons, and URLs, which are the features of tweets that have been usually discarded from analysis in previous research. And second, an approach to periodically identify influential users who are relevant to the sentiment towards a given Twitter user account. These methods were tested on the race for the 2017 Ecuadorian Presidential election.

Concerning the first approach, sentiment analyses were performed twice with the same dataset using a text analytical software tool called MeaningCloudTM: once as tweets were originally extracted, and again after pre-processing. The proposed pre-processing approach involved converting hashtags and emoticons into readable words and determining the sentiment for the most shared URLs. Then, Twitter user accounts were categorised and aggregated according to which of the two candidates their tweets had been classified as positive towards. Twitter sentiment analysis results after the pre-processing performed better, in terms of prediction accuracy, sentiment processing time, and for avoiding system crashes, than when using raw tweets, and more precise than the conventional polling organisations. Lastly, the volume of

unique Twitter users producing supportive tweets about their candidates provided a realistic measure of vote share.

Regarding the second approach, it developed from the observation of the network distribution which found that 80% of the extracted tweets used for analysis were retweets, which suggests that in the case study Twitter functioned mainly as a dissemination tool. Based on this observation, the five Twitter user accounts with the highest number of retweets every week were defined as the most influential users, which accounted for almost 44% of the total number of retweets, and they were categorised according to their sentiment towards each of the candidates. The analysis revealed that in most cases, the most influential users had a positive sentiment towards the winning candidate and negative sentiment towards the losing one. This provided some empirical evidence about the ability of influential users to disseminate sentiment and furthered awareness of the importance of producing most retweeted content.

The third paper proposes a novel predictive model based on the ER rule to determine the impact of tweets based on their number of retweets. The paper shows that the MAKER-RIMER approach can help deal with data with uncertainty, as introduced in Section 6.1. In this study, uncertainty can arise from the fact that the datasets did not contain all possible combinations of input variables. What is new is that the MAKER-RIMER approach allows splitting a model into partial models, which aims at maximising the use of available data, and then combining them together. Outside this case, MAKER-RIMER offers an alternative machine learning approach when datasets have limited data availability.

Using the same case study of the 2017 Ecuadorian Presidential election, the tweets of the two most voted candidates and the number of retweets that these obtained were used to build the predictive model. Besides MAKER-RIMER, other machine learning approaches, namely logistic regression, Naïve Bayes, decision tree, and support vector machine, were simultaneously tested for comparison purposes. The proposed MAKER-RIMER approach provided several advantages in comparison with the other machine learning methods, including, first, its suitability to analyse Twitter datasets based on likelihood data analysis. Second, the MAKER-RIMER approach overcomes the limitation of data availability by combining variables with the highest number of records available. And third, for the case study used in this

research, the MAKER-RIMER model performed better in that it has shown a smaller MCE than the other machine learning methods.

7.2. Contributions and implications

This thesis has proposed a set of alternative approaches to modelling and analysis of Twitter data, which can be of interest to researchers and campaigners of political parties. As emphasised in Chapter 4, the literature addressing Twitter as source of data for prediction algorithms is inconclusive. In this regard, the key contributions that this thesis has made to the literature and their implications for methodology and practice are summarised below:

- a) Identification of some of the theoretical and empirical research that is necessary to further develop the ability of Twitter content to understand and analyse user behaviour.
- b) An approach to pre-process Twitter data for sentiment analysis. As shown in Chapter 5, by coding and integrating hashtags, emoticons, and URLs in the sentiment analysis, the analytical procedure was more efficient, and the accuracy of prediction surged upwards. This constitutes a methodological contribution with practical implications for both the ability to measure and anticipate outcomes, and analysis processing time.
- c) An alternative two-stage approach to periodically identify and rank influential users that are relevant to the sentiment towards candidates during electoral campaigns. Unlike previous research, in which the definition of influential users is based on pre-established conceptualisations, the definition used in this study has emerged from the data. This means that influence can occur through different mechanisms and is measured in different ways. Thus, identifying the most prevalent mechanism of attempting influence is relevant to defining influential users.
- d) The finding that the number of Twitter users producing positive sentiment towards a candidate provides an effective way to measure political outcomes. This constitutes an empirical contribution with implications for political marketing campaigning in that it provides a realistic picture of user preferences during electoral campaigns.
- e) A novel predictive model based on the ER rule to predict the impact of a tweet. MAKER-RIMER splits a model into partial models and combines them back

together. Despite the uncertainty in the datasets used in this study, which do not allow all possible combinations of input variables, the proposed MAKER-RIMER approach has performed well since it has shown lower levels of MCE in predicting the impact of tweets than other machine learning approaches. Lastly, the results of MAKER-RIMER are easier to interpret. In this study, this means that the model allowed identifying what tweet features lead to high/low impact.

7.3. Reflections on strengths and limitations

This thesis has presented some strengths and limitations. A key strength lies in the case study used in this work. The high polarisation between partisans of opposing candidates of the 2017 Ecuadorian Presidential election may have been convenient for conducting sentiment analysis and developing predictive models. As for limitations, the scope of this study was limited to one case study to develop and validate the models. This limitation emerged because of the time spent in the current case study, the initial struggles to process the data, and the difficulty of identifying other appropriate case studies within the timescale to complete the PhD programme. Similarly, the likely presence of fake accounts or misleading content was not addressed, although it is sensible to assume their influence on the sentiment and the predictive model. Another limitation came from the fact that software for sentiment analysis in non-English languages is still developing. This means that the sentiment corpora used for this work (in the Spanish language) are more prone to error than for similar studies conducted in English. Lastly, it should be pointed out that the predictors used in this thesis are not exhaustive. This means that additional or different variables, as well as different ways to pre-process data, may affect the performance of the models presented in this study.

7.4. Directions for future research

This thesis opens up an agenda for future research in the following directions. First, there is a need and an opportunity to develop guiding principles for sampling Twitter data intended for modelling and understanding user behaviour which address issues of intuition and bias. Second, although this thesis has introduced an approach to coding and integrating the knowledge lying in hashtags, emoticons, and URLs in sentiment and predictive models, future work should look at algorithms that optimise these processes. Third, the growing literature on fake accounts and misleading content

in social media has focused by and large on the identification of veracity, which on, its own, is of limited relevance for predictive modelling. In this regard, future work should examine how fake accounts and their content impact user choice and behaviour. Fourth, future work should also provide guidelines to identify users influencing Twitter sentiment in a systematic and dynamic manner. Finally, future work should continue testing the MAKER-RIMER model in different social media platforms and competitive circumstances, and with different types of input variables.

References

- Al Hamoud, A., Alwehaibi, A., Roy, K., & Bikdash, M. (2018). Classifying political tweets using Naïve Bayes and support vector machine. In *31st International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems* (pp. 736-744). Springer, Montreal, Quebec, Canada.
- Alayba, A. M., Palade, V., England, M., & Iqbal, R. (2017). Arabic language sentiment analysis on health services. In *First International Workshop on Arabic Script Analysis and Recognition* (pp. 114-118). IEEE, Nancy, France.
- Bao, Y., Quan, C., Wang, L., & Ren, F. (2014). The role of pre-processing in Twitter sentiment analysis. In *Tenth International Conference on Intelligent Computing Methodologies* (pp. 615-624). Springer, Taiyuan, China.
- Bertrand, W. M., & Fransoo, J. C. (2002). Operations management research methodologies using quantitative modeling. *International Journal of Operations & Production Management*, 22(2), 241-264.
- Brownlee, J. (2017). Difference between classification and regression in machine learning. Retrieved on July 24th, 2019 from <https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/>.
- Ceron, A., Curini, L., & Iacus, S. M. (2015). Using sentiment analysis to monitor electoral campaigns: Method matters - Evidence from the United States and Italy. *Journal of Social Science Computer Review*, 33(1), 3-20.
- Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring user influence in Twitter: The million follower fallacy. In *Proceedings of the Fourth International AAAI Conference on Web and Social Media* (pp. 10-18). AAAI Press, Menlo Park, California, US.
- Choi, S. (2014). Flow, diversity, form, and influence of political talk in social-media-based public forums. *Journal of Human Communication Research*, 40(2), 209-237.
- Clement, J. (2019). Number of Twitter users worldwide from 2014 to 2020 (in millions). Retrieved on September 2nd, 2019 from <https://www.statista.com/statistics/303681/twitter-users-worldwide/>.
- de Maertelaere, M., Li, T., & Berens, G. (2012). Social influence: The effect of Twitter information on corporate image. In *Proceedings of the 14th Annual International Conference on Electronic Commerce* (pp. 292-293). ACM, Singapore.
- DiFranzo, D., & Gloria-Garcia, K. (2017). Filter bubbles and fake news. *Journal of ACM Crossroads*, 23(3), 32-48.
- Elghazaly, T., Mahmoud, A., & Hefny, H. A. (2016). Political sentiment analysis using Twitter data. In *Proceedings of the 16th International Conference on Internet of Things and Cloud Computing* (pp. 1-5). ACM, Cambridge, UK.
- Gabriel, D. (2013). Inductive and deductive approaches to research. Retrieved on August 9th, 2019 from <https://deborahgabriel.com/2013/03/17/inductive-and-deductive-approaches-to-research/>.
- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137-144.
- Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N., & Perera, A. (2012). Opinion mining and sentiment analysis on a Twitter data stream. In *International Conference on Advances in ICT for Emerging Regions* (pp. 182-188). IEEE, Colombo, Sri Lanka.

- Gupta, A., Lamba, H., & Kumaraguru, P. (2013). \$1.00 per RT #BostonMarathon #PrayForBoston: Analyzing fake content on Twitter. In *Proceedings of the 2013 eCrime Researchers Summit* (pp. 1-12). IEEE, Delhi, India.
- Gurajala, S., White, J. S., Hudson, B., & Matthews, J. N. (2015). Fake Twitter accounts: Profile characteristics obtained using an activity-based pattern detection approach. In *Proceedings of the 2015 International Conference on Social Media & Society* (pp. 1-7). ACM, Toronto, Canada.
- Hadjerrouit, S. (2008). Technology-enhanced learning in school algebra: The case of Aplusix. In *International Conference on Society for Information Technology & Teacher Education* (pp. 4453-4460). Association for the Advancement of Computing in Education, Vancouver, Canada.
- Hong, L., Dan, O., & Davison, B. D. (2011). Predicting popular messages in Twitter. In *Proceedings of the 20th International Conference Companion on World Wide Web* (pp. 57-58). ACM, Hyderabad, India.
- Hürlimann, M., Davis, B., Cortis, K., Freitas, A., Handschuh, S., & Fernández, S. (2016). A Twitter sentiment gold standard for the Brexit referendum. In *Proceedings of the 12th International Conference on Semantic Systems* (pp. 193-196). ACM, Leipzig, Germany.
- Ingle, A., Kante, A., Samak, S., & Kumari, A. (2015). Sentiment analysis of Twitter data using Hadoop. *International Journal of Engineering Research and General Science*, 3(6), 144-147.
- Isson, J. P. (2018). *Unstructured data analytics: How to improve customer acquisition, customer retention, and fraud detection and prevention*: John Wiley & Sons.
- Jain, A., & Jain, V. (2019). Sentiment classification of Twitter data belonging to renewable energy using machine learning. *Journal of Information and Optimization Sciences*, 40(2), 521-533.
- Jull, A., Bermingham, A., Adeosun, A., Ni Mhurchú, C., & Smeaton, A. F. (2016). Using Twitter for public health infoveillance: A feasibility study. In *Proceedings of the Second Twitter for Research Conference* (pp. 1-5). INSIGHT Centre for Data Analytics, Galway, Ireland.
- Khatri, S. K., & Srivastava, A. (2016). Using sentimental analysis in prediction of stock market investment. In *5th International Conference on Reliability, Infocom Technologies, and Optimization* (pp. 566-569). IEEE, Noida, India.
- Kumar, C. P., & Babu, L. D. (2019). Novel text pre-processing framework for sentiment analysis. *Smart Intelligent Computing and Applications* (pp. 309-317): Springer.
- Montangero, M., & Furini, M. (2015). Trank: Ranking Twitter users according to specific topics. In *12th Annual Conference on Consumer Communications and Networking* (pp. 767-772). IEEE, Las Vegas, Nevada, US.
- Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., & Stoyanov, V. (2016). SemEval-2016 task 4: Sentiment analysis in Twitter. *arXiv preprint arXiv:1912.01973*.
- Naveed, N., Gottron, T., Kunegis, J., & Alhadi, A. C. (2011). Bad news travel fast: A content-based analysis of interestingness on Twitter. In *Proceedings of the 3rd International Web Science Conference* (pp. 1-7). ACM, Koblenz, Germany.
- Nesi, P., Pantaleo, G., Paoli, I., & Zaza, I. (2018). Assessing the retweet proneness of tweets: Predictive models for retweeting. *Journal of Multimedia Tools and Applications*, 77(20), 26371–26396.
- Oliveira, N., Cortez, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading

- volume and survey sentiment indices. *Journal of Expert Systems with Applications*, 73, 125-144.
- Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016). Sentiment analysis of Twitter data for predicting stock market movements. In *International Conference on Signal Processing, Communication, Power and Embedded System* (pp. 1345-1350). IEEE, Paralakhemundi, India.
- Pandarachalil, R., Sendhilkumar, S., & Mahalakshmi, G. (2015). Twitter sentiment analysis for large-scale data: An unsupervised approach. *Journal of Cognitive Computation*, 7(2), 254-262.
- Petrovic, S., Osborne, M., & Lavrenko, V. (2011). RT to win! Predicting message propagation in Twitter. In *Proceedings of the Fifth International AAAI Conference on Web and Social Media* (pp. 586-589). AAAI Press, Barcelona, Spain.
- Pibernik, J., & Dolić, J. (2008). A design-based research framework for assessing e-learning in sustainable development. In *10th International Design Conference on Design of Graphic Media* (pp. 1433-1438). University of Zagreb, Faculty of Graphic Arts, Croatia, Dubrovnik, Croatia.
- R Core Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (Version 1.0.136). Retrieved from <http://www.r-project.org/>.
- Schlegel, G. L. (2014). Utilizing Big data and predictive analytics to manage supply chain risk. *The Journal of Business Forecasting*, 33(4), 11-17.
- Sharma, P., & Moh, T.-S. (2016). Prediction of Indian election using sentiment analysis on Hindi Twitter. In *International Conference on Big Data* (pp. 1966-1971). IEEE, Washington, US.
- Shin, J., Jian, L., Driscoll, K., & Bar, F. (2018). The diffusion of misinformation on social media: Temporal pattern, message, and source. *Journal of Computers in Human Behavior*, 83, 278-287.
- Sinha, S., Dyer, C., Gimpel, K., & Smith, N. A. (2013). Predicting the NFL using Twitter. *arXiv preprint arXiv:1310.6998*.
- Skuzza, M., & Romanowski, A. (2015). Sentiment analysis of Twitter data within Big data distributed environment for stock prediction. In *Federated Conference on Computer Science and Information Systems* (pp. 1349-1354). IEEE, Lodz, Poland.
- Thakkar, H., & Patel, D. (2015). Approaches for sentiment analysis on Twitter: A state-of-art study. *arXiv preprint arXiv:1512.01043*.
- Trilling, D., Tolochko, P., & Burscher, B. (2017). From newsworthiness to shareworthiness: How to predict news sharing based on article characteristics? *Journalism & Mass Communication Quarterly*, 94(1), 38-60.
- Vijayaraghavan, P., Vosoughi, S., & Roy, D. (2017). Twitter demographic classification using deep multi-modal multi-task learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (pp. 478-483). Association for Computational Linguistics, Vancouver, Canada.
- Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and Big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics*, 34(2), 77-84.
- Wright, L. A., Golder, S., Balkham, A., & McCambridge, J. (2019). Understanding public opinion to the introduction of minimum unit pricing in Scotland: A qualitative study using Twitter. *Journal of BMJ Open*, 9(6), 1-8.

Yu, Y., & Wang, X. (2015). World Cup 2014 in the Twitter world: A Big data analysis of sentiments in US sports fans' tweets. *Journal of Computers in Human Behavior*, 48, 392-400.