News sentiment impact analysis (NSIA) framework

Author:
Al-Qudah, Islam

Publication Date:
2018

DOI:
https://doi.org/10.26190/unsworks/20210

License:
https://creativecommons.org/licenses/by-nc-nd/3.0/au/
Link to license to see what you are allowed to do with this resource.

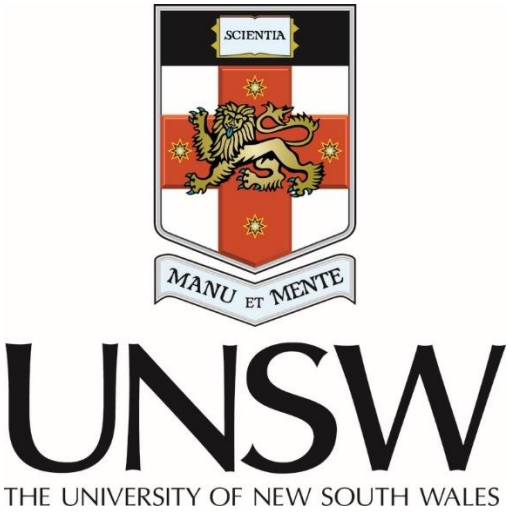
Downloaded from http://hdl.handle.net/1959.4/59504 in https://unsworks.unsw.edu.au on 2024-04-25

A thesis in fulfilment of the requirements for the degree of

Doctor of Philosophy

NEWS SENTIMENT IMPACT ANALYSIS (NSIA) FRAMEWORK

Islam Khalid Yousif AL-QUDAH



School of Computer Science and Engineering

Faculty of Engineering

January 2018

Surname or Family name: **AL-QUDAH**
First name: **Islam** Other name/s: **Khalid Yousif**
Abbreviation for degree as given in the University calendar: **PhD**
School: **School of Computer Science and Engineering** Faculty: **Faculty of Engineering**
Title: **News Sentiment Impact Analysis (NSIA) Framework**

Abstract 350 words maximum:

The huge increase in online content has rapidly stepped up the challenges facing various entities such as organisations, businesses, governments. Dealing with this immense increase in information is beyond any human ability, as it requires massive efforts to go through huge volumes of posts/news to analyse and understand their content. In responding to this challenge, research has picked up in recent years to automatically analyse, and extract opinions/sentiment from online content.

This thesis investigated the area of sentiment analysis and its impact on financial market entities. The thesis includes a review of the sentiment analysis processes, , an overview of sentiment analysis studies and their application to the financial markets domain. The literature shows a gap in defining systematic and reusable evaluation processes, for users to automatically conduct impact analysis of sentiment datasets in different financial contexts. In addressing the research gap the thesis proposes a framework called News Sentiment Impact Analysis (NSIA) which consists of three components which are a conceptual data model, a software architecture, and a set of use cases. The key component which is the conceptual data model captures three sets of parameters: the financial context parameters, the sentiment related parameters, and the impact measure parameters. The CPD model is supported by a software architecture, which consists of a GUI, Business, and Data layers. The main use case allows the analyst to define the CPD parameters using the provided GUI layer and trigger the impact analysis process. To evaluate the proposed framework, a prototype is implemented. Three case studies have been conducted to evaluate the different aspects of the proposed framework.

The case studies results show that the proposed NSIA framework meets the research objectives and demonstrate the flexibility of the proposed NSIA framework for conducting impact analysis in various contexts, using different sentiment extractions techniques, and various impact measures. The results also show that the framework is able to support a systematic methodology that enables reproducibility and consistency in conducting impact analysis studies. Moreover, the evaluation was conducted using the framework's architecture and prototype which enables automation of the impact analysis studies, and eliminates the possibility of human errors.

Declaration relating to disposition of project thesis/dissertation

I hereby grant to the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or in part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all property rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstracts International (this is applicable to doctoral theses only).

.....	.....	.....
Signature	Witness Signature	Date

The University recognises that there may be exceptional circumstances requiring restrictions on copying or conditions on use. Requests for restriction for a period of up to 2 years must be made in writing. Requests for a longer period of restriction may be considered in exceptional circumstances and require the approval of the Dean of Graduate Research.

FOR OFFICE USE ONLY Date of completion of requirements for Award:

ORIGINALITY STATEMENT

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'

Signed ……..

Date ……..

COPYRIGHT STATEMENT

'I hereby grant the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all proprietary rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.
I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstract International (this is applicable to doctoral theses only).
I have either used no substantial portions of copyright material in my thesis or I have obtained permission to use copyright material; where permission has not been granted I have applied/will apply for a partial restriction of the digital copy of my thesis or dissertation.'

Signed …….....

Date …….....

AUTHENTICITY STATEMENT

'I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis. No emendation of content has occurred and if there are any minor variations in formatting, they are the result of the conversion to digital format.'

Signed …….....

Date …….....

CONTENTS

ABSTRACT ..... **IX**
ACKNOWLEDGMENTS ..... **X**
LIST OF PUBLICATIONS ..... **XI**
LIST OF FIGURES ..... **XII**
LIST OF TABLES ..... **XIV**
LIST OF ABBREVIATIONS ..... **XVI**
1 INTRODUCTION **1**
1.1 BACKGROUND 1
1.2 APPLYING SENTIMENT ANALYSIS TO THE FINANCIAL MARKETS DOMAIN 2
1.3 PROBLEM STATEMENT AND THESIS OBJECTIVES 3
1.4 THESIS STRUCTURE 3
2 LITERATURE REVIEW **5**
2.1 TYPES AND SOURCES OF TEXT CORPUS 5
2.2 SENTIMENT ANALYSIS PROCESSES 6
2.2.1 Acquire Text Corpus (ATC) *7*
2.2.2 Text Pre-Processing (TPP) *8*
2.2.3 Calculate Sentiment Metrics (CSM) *10*
2.2.4 News analytics datasets *12*
2.2.5 Sentiment analysis accuracy metrics *13*
2.3 OVERVIEW OF FINANCIAL MARKETS 14
2.3.1 Financial markets definitions *14*
2.3.2 Financial markets data *15*
2.3.3 Market measures *16*
2.4 SENTIMENT ANALYSIS IN FINANCE RESEARCH 17
2.4.1 Sentiment analysis techniques used in finance studies *18*

2.4.2 Techniques used to evaluate impact *23*

2.4.3 Discussion *30*

2.5 CONCLUSION 31

3 RESEARCH METHODOLOGY 33

3.1 RESEARCH PROBLEM 33

3.2 RESEARCH QUESTIONS 33

3.3 RESEARCH APPROACH 34

3.4 RESEARCH PROCESS 34

3.5 CONCLUSION 36

4 NSIA FRAMEWORK 37

4.1 OVERVIEW OF THE NSIA FRAMEWORK 37

4.2 NSIA DATA MODEL 38

4.2.1 Market Data (MD) model *38*

4.2.2 Sentiment Data (SD) model *40*

4.2.3 Comparison Parameters Data (CPD) model *41*

4.3 NSIA ARCHITECTURE 47

4.3.1 Overview of NSIA architecture *47*

4.3.2 Business layer *48*

4.3.3 Data layer *49*

4.4 NSIA USE CASES 49

4.4.1 Overview of use cases *49*

4.4.2 Define financial context parameters *50*

4.4.3 Define sentiment extraction parameters *51*

4.4.4 Conduct impact analysis *52*

4.5 CONCLUSION 53

5 PROTOTYPE IMPLEMENTATION 55
5.1 IMPLEMENTING THE DATA LAYER 55
5.1.1 Datasets used 55
5.1.2 Implementation Choices 56
5.1.3 Implementing the Market Data (MD) model 57
5.1.4 Implementing the Sentiment Data (SD) model 60
5.1.5 Implementing the Comparison Parameters Data (CPD) model 60
5.2 IMPLEMENTING THE BUSINESS LAYER 62
5.2.1 Implementation choices 62
5.2.2 Implementing the Sentiment Processing (SP) component 63
5.2.3 Implementing Impact Analysis Component 66
5.3 IMPLEMENTING THE GUI LAYER 68
5.3.1 Define the FC parameters 68
5.3.2 Define the sentiment extraction parameters 69
5.3.3 Conduct Impact Analysis use case 69
5.4 LIMITATIONS 69
5.5 CONCLUSION 70
6 NSIA FRAMEWORK EVALUATION 71
6.1 NSIA FRAMEWORK EVALUATION 71
6.2 CASE STUDY 1: NEGATIVE NEWS DAILY IMPACT ON TWO COMPANIES 72
6.2.1 Defining the CPD model parameters 72
6.2.2 Performing the use cases 74
6.2.3 Results discussion 75
6.2.4 Discussion 77
6.3 CASE STUDY2: NEGATIVE NEWS DAILY IMPACT IN MULTIPLE CONTEXTS 77

6.3.1 Defining the CPD model parameters *78*

6.3.2 Performing the use cases *81*

6.3.3 Results discussion *83*

6.3.4 Discussion *86*

6.4 CASE STUDY3: NEGATIVE NEWS INTRADAY IMPACT 86

6.4.1 Case study scenario *86*

6.4.2 Defining the CPD model parameters *87*

6.4.3 Performing the use cases *89*

6.4.4 Results discussion *90*

6.5 DISCUSSION 95

6.6 CONCLUSION 98

7 CONCLUSION AND FUTURE WORK 99

7.1 THESIS SUMMARY 99

7.2 ADDRESSING THE RESEARCH QUESTIONS 100

7.3 BENEFITS OF THE RESEARCH 101

7.4 THESIS LIMITATIONS 102

7.5 FUTURE WORK 103

8 REFERENCES 105

APPENDIX A: ADDITIONAL INFORMATION FOR NSIA FRAMEWORK .. 117

APPENDIX B: GUI IMPLEMENTATION OF NSIA FRAMEWORK 120

APPENDIX C: *EXTREME_NEWS_ALGORITHM* PSEUDO CODE 123

ABSTRACT

The huge increase in online content has rapidly stepped up the challenges facing various entities such as organisations, businesses, governments. Dealing with this immense increase in information is beyond any human ability, as it requires massive efforts to go through huge volumes of posts/news to analyse and understand their content. In responding to this challenge, research has picked up in recent years to automatically analyse, and extract opinions/sentiment from online content.

This thesis investigated the area of sentiment analysis and its impact on financial market entities. The thesis includes a review of the sentiment analysis processes, an overview of sentiment analysis studies and their application to the financial markets domain. The literature shows a gap in defining systematic and reusable evaluation processes, for users to automatically conduct impact analysis of sentiment datasets in different financial contexts. In addressing the research gap the thesis proposes a framework called News Sentiment Impact Analysis (NSIA) which consists of three components which are a conceptual data model, a software architecture, and a set of use cases. The key component which is the conceptual data model captures three sets of parameters: the financial context parameters, the sentiment related parameters, and the impact measure parameters. The CPD model is supported by a software architecture, which consists of a GUI, Business, and Data layers. The main use case allows the analyst to define the CPD parameters using the provided GUI layer and trigger the impact analysis process. To evaluate the proposed framework, a prototype is implemented. Three case studies have been conducted to evaluate the different aspects of the proposed framework.

The case studies results show that the proposed NSIA framework meets the research objectives and demonstrate the flexibility of the proposed NSIA framework for conducting impact analysis in various contexts, using different sentiment extractions techniques, and various impact measures. The results also show that the framework is able to support a systematic methodology that enables reproducibility and consistency in conducting impact analysis studies. Moreover, the evaluation was conducted using the framework's architecture and prototype which enables automation of the impact analysis studies, and eliminates the possibility of human errors.

ACKNOWLEDGMENTS

All praises are due to Allah (S.W.T) and to Him alone. For all the bounties, He bestowed on me to enable me to finish this thesis.

My appreciation goes to many people around me who helped me to complete this thesis. First and foremost, I would like to thank my supervisor Professor Fethi Rabhi for his time, effort, careful advice and timely encouragement throughout the last four years, which has helped me significantly to develop my research skills. I would like to extend my gratitude to Associate Professor Maurice Peat from the University of Sydney for his great advice and suggestions to improve the quality of the work in this thesis. Special thanks go to Associate Professor Brahim Saadouni from the University of Manchester for his great comments and advice while conducting the case studies. I would like also to thank Sirca for providing the financial market data and the sentiment data used in the case studies.

It won't have been possible to accomplish the thesis without the continuous support of my beloved wife Taima. Her patience, tolerance and understanding during the last four years has made my work on the thesis much easier. I would like to extend my thanks to my two beautiful daughters Sarah (10 years old) and Leen (4 years old) for their love and support.

Last but not least, I would like to thank my parents for their continuous love, support and encouragement throughout these years of my PhD study.

LIST OF PUBLICATIONS

Below is a list of relevant publications in which the author has authored.

1- Qudah, I., Rabhi, F. A., & Peat, M. (2014). A proposed framework for evaluating the effectiveness of financial news sentiment scoring datasets. *Enterprise applications and services in the finance industry* (pp. 29-47) Springer.

2- Qudah, I., & Rabhi, F. A. (2016). News sentiment impact analysis (NSIA) framework. *International Workshop on Enterprise Applications and Services in the Finance Industry*, pp. 1-16.

LIST OF FIGURES

FIGURE 2.1 TEXT CORPUS SOURCES 6

FIGURE 2.2 SENTIMENT ANALYSIS PROCESSES 7

FIGURE 2.3 TEXT PRE-PROCESSING 8

FIGURE 2.4 EXAMPLE OF SENTICNET USING ONTOLOGY BASED APPROACH (SENTICNET, 2014) 11

FIGURE 3.1 RESEARCH STAGES IN RESEARCH PROCESS 35

FIGURE 4.1 NSIA FRAMEWORK COMPONENTS 37

FIGURE 4.2 NSIA DATA MODEL COMPONENTS 38

FIGURE 4.3 MARKET DATA CONCEPTUAL MODEL 39

FIGURE 4.4 SENTIMENT DATA CONCEPTUAL MODEL 40

FIGURE 4.5 HIGH LEVEL VIEW OF CPD MODEL 43

FIGURE 4.6 DEFINING FINANCIAL CONTEXT MODEL 43

FIGURE 4.7 DEFINING SN PARAMETER ENTITIES 44

FIGURE 4.8 DEFINING IMPACT MEASURES PARAMETERS 45

FIGURE 4.9 NSIA ARCHITECTURE 48

FIGURE 4.10 USE CASE TO DEFINE CPD PARAMETERS AND EXECUTE IMPACT ANALYSIS STUDIES 50

FIGURE 4.11 DEFINE FINANCIAL CONTEXT PARAMETERS SEQUENCE DIAGRAM 51

FIGURE 4.12 DEFINE SENTIMENT EXTRACTION (SN) PARAMETERS SEQUENCE DIAGRAM 52

FIGURE 4.13 CONDUCT IMPACT ANALYSIS SEQUENCE DIAGRAM 53

FIGURE 5.1 THOMSON REUTERS TICK HISTORY WEB PORTAL SHOWING TRADES AND QUOTES 56

FIGURE 5.2 THOMSON REUTERS TICK HISTORY WEB PORTAL SHOWING MARKET DEPTH DATA 56

FIGURE 5.3 ORACLE SQL DEVELOPER IMPORT DATA UTILITY (ORACLE CORPORATION, 2017A) 57

FIGURE 5.4 CREATES INTRADAY_HOMO_TS TABLE STRUCTURE 59

FIGURE 5.5 SAMPLE INTRADAY RETURNS FOR GERMAN COMPANY SIEMENS 60
FIGURE 5.6 MAPPING SD ENTITIES TO SENT_RAW_DATA DATABASE OBJECT 60
FIGURE 5.7 FILTRATION_FUNCTION METHOD DEFINITION 63
FIGURE 5.8 EXTREME_NEWS_ALGORITHM ALGORITHM DEFINITION 65
FIGURE 6.1 SUM OF ALL MCARS FOR THE 24 EXPERIMENTS BY ESE ALGORITHMS 85
FIGURE 6.2 IMPACT GROUPED BY COUNTRY AND EXTREME SENTIMENT EXTRACTION ALGORITHMS ... 86
FIGURE 6.3 INTRADAY MCAAR AND LBM RESULTS FOR STUDY NO. 3 95
FIGURE 6.4 INTRADAY MCAAR AND PRICE JUMPS STATISTICS RESULTS FOR STUDY NO. 3 95

LIST OF TABLES

TABLE 2.1 PRE-PROCESSING STEP USED IN A SAMPLE OF SENTIMENT ANALYSIS STUDIES 9

TABLE 2.2 UNSCHEDULED NEWS SOURCES 20

TABLE 2.3 SCHEDULED NEWS SOURCES 21

TABLE 2.4 SENTIMENT ANALYSIS APPROACHES AND IMPACT MODELS 24

TABLE 4.1 FINANCIAL CONTEXT PARAMETERS ($\boldsymbol{FC}$) 41

TABLE 4.2 SENTIMENT EXTRACTION PARAMETERS ($\boldsymbol{SN}$) 42

TABLE 4.3 IMPACT MEASURE PARAMETERS ($\boldsymbol{IM}$) 42

TABLE 5.1 MAPPING THE MARKET DATA MODEL ENTITIES TO DATABASE OBJECTS 58

TABLE 5.2 HGA ROLE IN PRODUCING TIMESERIES CSV FILES 59

TABLE 5.3 MAPPING CPD OBJECTS AND THE IMPLEMENTATION OBJECTS 61

TABLE 5.4 MAPPING TECHNOLOGIES TO THEIR CORRESPONDING IMPLEMENTED IMPACT MODELS 62

TABLE 5.5 FILTRATION ATTRIBUTES IN TRNA DATASET 63

TABLE 5.6 FILTRATION FUNCTIONS (FA) WITH FILTRATION ATTRIBUTES 64

TABLE 5.7 PROTOTYPE EXTREME_NEWS_ALGORITHM ALGORITHMS 65

TABLE 5.8 TOOLS USED IN IMPLEMENTING THE IMPACT MODELS 66

TABLE 6.1 RELATING CASE STUDIES AND RESEARCH OBJECTIVES 72

TABLE 6.2 DEFINING FINANCIAL CONTEXT (FC) PARAMETERS 73

TABLE 6.3 DEFINING THE SN PARAMETERS 73

TABLE 6.4 DEFINING THE IM PARAMETERS 74

TABLE 6.5 DIFFERENT CPD PARAMETERS USED IN CASE STUDY 1 75

TABLE 6.6 IMPACT STUDIES RESULTS 76

TABLE 6.7 FINANCIAL CONTEXT ENTITIES 78
TABLE 6.8 DEFINING FINANCIAL CONTEXT (FC) PARAMETERS 79
TABLE 6.9 DEFINING THE SN PARAMETERS 80
TABLE 6.10 DEFINING THE IM PARAMETERS 80
TABLE 6.11 DIFFERENT CPD PARAMETERS USED IN CASE STUDY 2 81
TABLE 6.12 IMPACT STUDIES RESULTS 83
TABLE 6.13 DEFINING FINANCIAL CONTEXT (FC) PARAMETERS 87
TABLE 6.14 DEFINING THE SN PARAMETERS 88
TABLE 6.15 DEFINING THE IM PARAMETERS 88
TABLE 6.16 DIFFERENT CPD PARAMETERS USED IN CASE STUDY 3 89
TABLE 6.17 INTRADAY MCAAR IMPACT RESULTS 91
TABLE 6.18 INTRADAY LBM IMPACT RESULTS 92
TABLE 6.19 INTRADAY PRICE JUMPS STATISTICS RESULTS 94
TABLE 6.20 INTRADAY VS DAILY MCAAR RESULTS 96

LIST OF ABBREVIATIONS

AMEX	American Stock Exchange
AORD	Australia's ASX All Ordinaries Index
API	Application Program Interfaces
ARCH	Auto Regressive Conditional Heteroskedasticity
ATC	Acquire Text Corpus
ATLI	Australia's ASX top 20 leaders
CPD	Comparison Parameters Data
CRSP	Center for Research in Securities Prices
CSM	Calculate Sentiment Metrics
CSV	Comma Separated Values
CXKNX	Germany's Industrial Index
DJI	Dow Jones Index
DJIA	Dow Jones Industrial Average
DMM	Data Model Management
EMH	Efficient Market Hypothesis
ESE	Extreme Sentiment Extraction
FC	Financial Context
GARCH	Generalized Auto Regressive Conditional Heteroskedasticity
GDAXI	Germany's DAX Index
GTSX	Canada's healthcare index
GUI	Graphical User Interface
HWI	Hardware Index in the USA

IA	Impact Analysis
IM	Impact Measure
LBM	Liquidity Based Model
LnR	Linear Regression
LoR	Logistic Regression
MCAAR	Mean Cumulative Average Abnormal Returns
MD	Market Data
ML	Machine Learning
MSH	Morgan Stanley High-Tech
NASDAQ	National Association of Securities Dealers Automated Quotations
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NSIA	News Sentiment Impact Analysis
NYSE	New York Stock Exchange
OLS	Ordinary Least Squares
POS	Part of Speech Tagging
REST	Representational State Transfer
RIC	Reuters Instrument Code
RPC	Remote Procedure Call
S&P	Standard and Poor
SD	Sentiment Data
SN	Sentiment Extraction
SP	Sentiment Processing
SPTSE	S&P Toronto Stock Exchange

TPP	Text Pre-Processing
TRNA	Thomson Reuters News Analytics
TRTH	Thomson Reuters Tick History
XML	eXtensible Markup Language

1 INTRODUCTION

This chapter introduces the research area addressed in this thesis. First, it gives some background about sentiment analysis in general, and its applicability to the field of finance in section 1.1. Section 1.2 gives a brief overview of sentiment analysis and its application to the domain of financial markets. In section 1.3, the thesis objectives are stated, and finally the thesis structure is outlined in section 1.4.

1.1 Background

In todays' fast growing online trade and commercial activity, the internet has become the enabling medium of both businesses and consumers. With the great opportunities it has delivered, comes new challenges. Customers usually express their opinions via social media platforms about the different services and products they use, therefore, it is becoming crucial for organizations and businesses around the world to capture and filter this huge influx of information. This is indeed a daunting task that is beyond any human ability. The problem with manually analysing opinions is that it requires massive efforts to go through thousands of posts/news items and analyse them. In addition, news are scattered over different heterogeneous sources (blogs, message boards, social networks, newswires), which makes the task of aggregating, analysing and extracting opinions/sentiment immensely difficult and very time consuming. To deal with this new challenge a fairly new research topic stemming from text capturing and analysing has emerged, called opinion mining or sentiment analysis, which deals specifically with processing opinions found in any text source.

This field of research concerning sentiment or opinion mining has been rigorously studied for more than a decade and many researchers and organizations have produced various sentiment analysis techniques, analysing news across various domains, such as:

- Product reviews and movie reviews extracting and weighting subjective text from products and movies reviews have been investigated by a wide range of studies (Pang et al., 2002; Jebaseeli & Kirubakaran, 2012; Dhaoui et al., 2017).
- Political races where the relationship between people expressing their opinions using mediums such as Facebook and twitter and elections results has been investigated by many such as (Tumasjan et al., 2010).

- Market intelligence derived from online discussions helps organizations and companies get timely information related to brands, products and strategies (Glance et al., 2005).
- Health services were the focus of a number of studies, in which patients' reviews and comments made on online websites related to the health services they received have been extracted and weighted (Greaves et al., 2013).
- Social events are another domain where sentiment analysis have been applied, investigating peoples' opinions around important social events such as musical concerts (Zhou et al., 2013).
- Sports betting was the focus of some studies, where for example (Hong & Skiena, 2010) investigated the efficacy of building a betting strategy using sentiment analysis of peoples' opinions of competing teams.

Conducting sentiment analysis is effectively a domain specific problem, because the vocabularies, dictionaries and linguistic rules vary across domains. In this thesis, we are focusing on sentiment analysis in the context of financial markets. The motivation is that although there have been the focus of many studies in literature, there are many open questions and challenges related to investigating the complex relationships between different sources of text (news articles, corporate announcements, tweets, blogs and message boards) and the financial markets (Das & Chen, 2007; Kothari et al., 2009). This is discussed in more detail next.

1.2 Applying sentiment analysis to the financial markets domain

Financial markets reaction to news has been studied extensively, starting few decades ago (Niederhoffer, 1971). Researchers have diverged in their investigation approaches, some focusing on studying the investor sentiment beliefs about companies' cash flows and associated risk in investing in these companies, using the facts they have in hand (Baker & Wurgler, 2006), and how these beliefs impact their decision making. Other researchers focused on studying text sentiment, weighing the subjectivity elements found in text. Between investor and text based sentiment, this thesis is concerned with the latter, as it usually captures investors' subjectivity (such as in tweets and blogs) and the subjectivity found in more formal sources such as news articles and corporate disclosers.

Researchers utilise various sources of information such as corporate disclosers, tweets, Facebook posts, news articles, internet message boards and blogs. Different sentiment analysis approaches have

been employed to extract and determine the subjectivity weight of such text sources. These approaches are dictionary based approaches, machine learning approaches and natural language processing approaches. In addition, to evaluate the effectiveness of their proposed sentiment analysis models, researchers have utilised various financial markets impact models. Regression analysis methods were used in many studies, while trading strategies were trialled by others. With the large amount of text involved and the mixture of models been used, conducting such studies by finance researchers has become a data intensive activity which requires a mixture analytic modelling and software analysis skills.

1.3 Problem statement and thesis objectives

The problem addressed in this thesis is that existing studies are difficult to reproduce outside a specific context, i.e. the processes involved in determining a sentiment metrics and evaluating its impact are not systematically documented. In particular, there is a lack in the way impact analysis can be conducted in multiple contexts. This is a challenge because of the difficulty in defining precisely what a context means in relation to sentiment analysis.

Consequently, the objectives of the thesis are to address these challenges from an end user's i.e. finance researcher's perspective. The outcome should be a system that improves the end user's experience in the following aspects:

- Provide *flexibility* in enabling users to derive and compare between different studies in terms of parameters, or datasets used.
- Propose a systematic methodology that enables *reproducibility* and *consistency* in conducting sentiment-driven impact analysis studies.
- Design software tools to automate the methodology so that impact analysis can be conducted efficiently and without the risk of errors.

1.4 Thesis structure

The remainder of this PhD thesis is organized into the following chapters:

- *Chapter 2 Literature Review*: surveys existing literature covering sentiment analysis and its application to financial markets and identifies the gaps in this research area.

- *Chapter 3 Research Methodology*: discusses the research questions, objectives and the methodology used in carrying out this research.
- *Chapter 4* proposes a framework, namely News Sentiment Impact Analysis (NSIA) framework and it is three main components NSIA data model, architecture and use cases.
- *Chapter 5* describes an implementation of the proposed framework.
- *Chapter 6* covers three case studies devised to validate the proposed framework. Two case studies were conducted using low frequency financial data, and a more comprehensive one used high frequency financial data.
- Chapter 7 concludes this thesis.

2 LITERATURE REVIEW

The focus of this thesis is on reviewing sentiment analysis processes and their application in the financial markets domain. First, we overview text (news stories, blogs, tweets, Facebook posts…etc.) sources and types in section 2.1. Second, in section 2.2 we overview the sentiment analysis processes. Third, we provide an overview of the financial markets theories, data, and measures in section 2.3. Next, in section 2.4, provides reviews and discusses studies using sentiment analysis in finance before concluding the chapter in section 2.5.

2.1 Types and sources of text corpus

"Sentiment analysis, also called opinion mining, is the field of study that analyses people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes" as defined by (Liu, 2012). Martin (2014) defined sentiment analysis as *"The process of algorithmically identifying and categorizing opinions expressed in text to determine the writer's attitude toward the subject of the document (or post)"*.

There are a number of text sources which are been captured while surveying text mining and sentiment analysis literature. A set of text (unstructured) documents is often referred to as a *text corpus* (Nordquist, 2016).

Text corpuses (news stories, blogs, tweets, Facebook posts…etc.) are collected and distributed by a number of major players in the field of text aggregators and publishers. Agence France-Presse (2016), Associated press (2016), Business wire (2016) and PR newswire (2016) are examples of text (news stories in particular) providers. Thomson Reuters (2014), Bloomberg (2016), Ravenpack (2016), and Yahoo Finance (2016) are another set of text (news stories in particular) corpuses aggregators, which distribute content (to subscribed clients) gathered by other news providers. Figure 2.1 covers some of the most important text corpuses sources.

News are generally divided into two types: scheduled and unscheduled. Scheduled news are news announcements scheduled to be published on a pre-known date, such as governments' announcements of unemployment rates, interest rate and inflation rates (Investopedia, 2016a). Other examples of scheduled news are announcements made by corporates regarding their earnings, management board changes (Robertson et al., 2013). Unscheduled news are published as soon as new information becomes available such as news press (newspapers, TV, Radio and Internet). Blogs, message boards,

social networks like Facebook and Twitter also fall under this category (Mitra and Mitra, 2011; Twitter About, 2016).

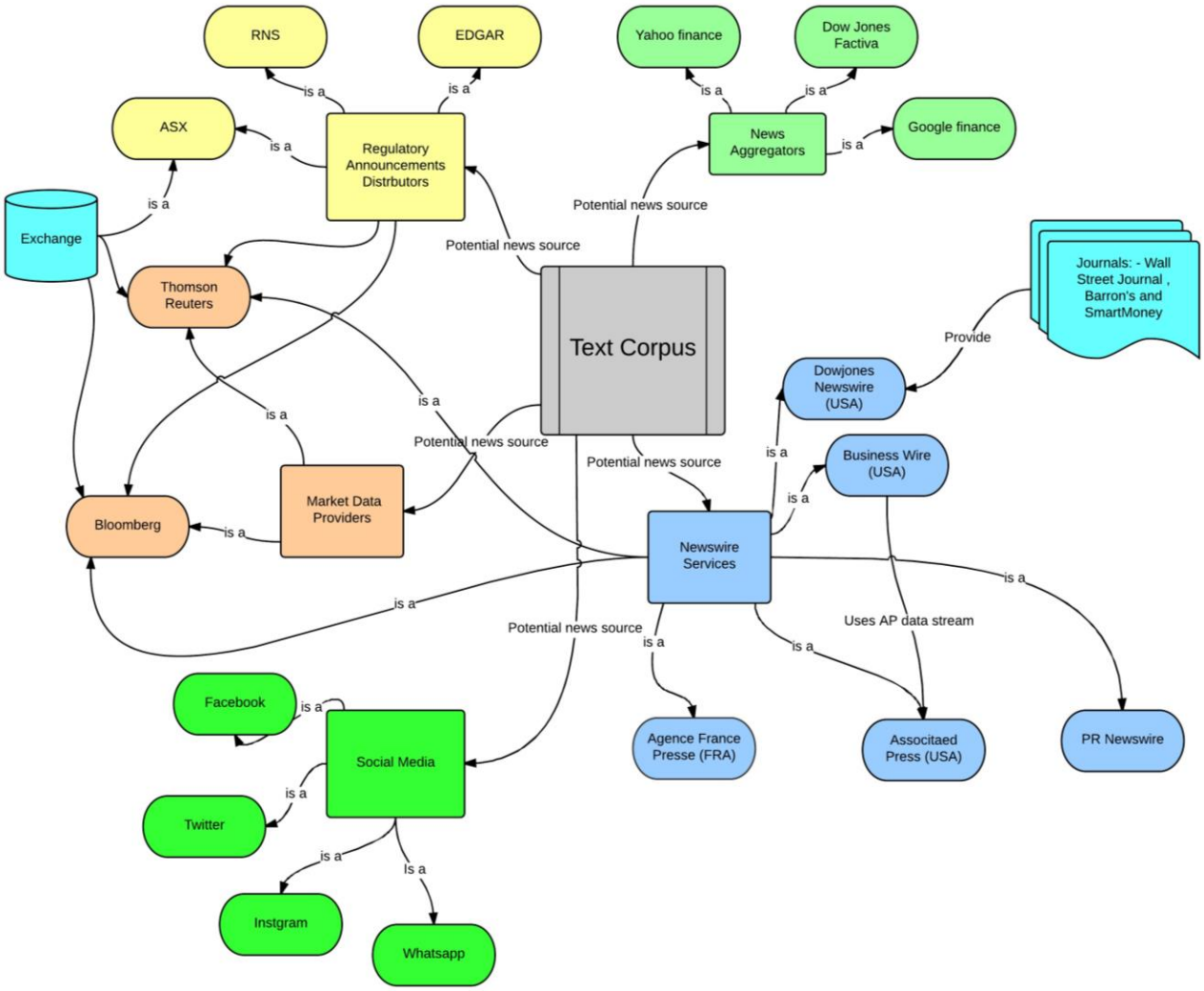


Figure 2.1 Text corpus sources

2.2 Sentiment analysis processes

Literature on sentiment analysis (e.g. Antweiler and Frank (2004); Das and Chen (2007); Liu (2012)) describes a number of common processes to analyse text and produce sentiment analytics metrics. These processes can be illustrated as: Acquire Text Corpus (ATC), Text Pre-Processing (TPP) and Calculate Sentiment Metrics (CSM) as featured in Figure 2.2. The goal of applying these processes

would be to generate a sentiment metrics dataset out of a text corpus. These processes are described next.

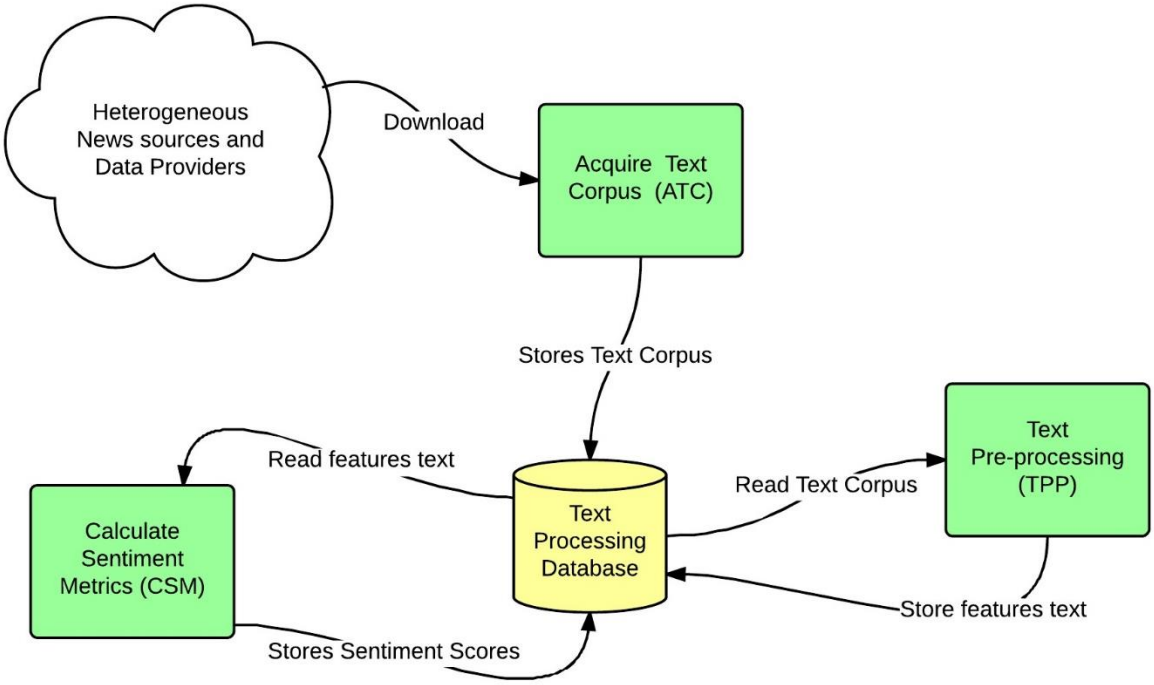


Figure 2.2 Sentiment analysis processes

2.2.1 Acquire Text Corpus (ATC)

Text corpuses can be aquired by clients (from text corpus providers) using a variety of techniques, which providers make available to their clients. Serveral known communication mediums are being used today to supply information such as: RSS feeds and Application Program Interfaces (API) requests (ProgrammableWeb, 2016). APIs are commonly used to connect clients with their providers, where different protocols are used such as: Remote Procedure Call (RPC) and Representational State Transfer (REST). The response is usually structured as JSON or as XML and returned to the client (RESTful, 2016; ProgrammableWeb, 2016). For example, Thomson Reuters and Bloomberg offer APIs to their vast pool of clients to get access to various text corpuses and/or financial markets data repositories (Thomson Reuters, 2014; Bloomberg, 2016).

2.2.2 Text Pre-Processing (TPP)

Text pre-processing aims at reducing acquired text corpuses into smaller and useful text features, by applying number of text cleansing and reduction techniques. The main common techniques found in literature are: tokenization, remove stop words, stemming, part of speech tagging, and finally features extraction.

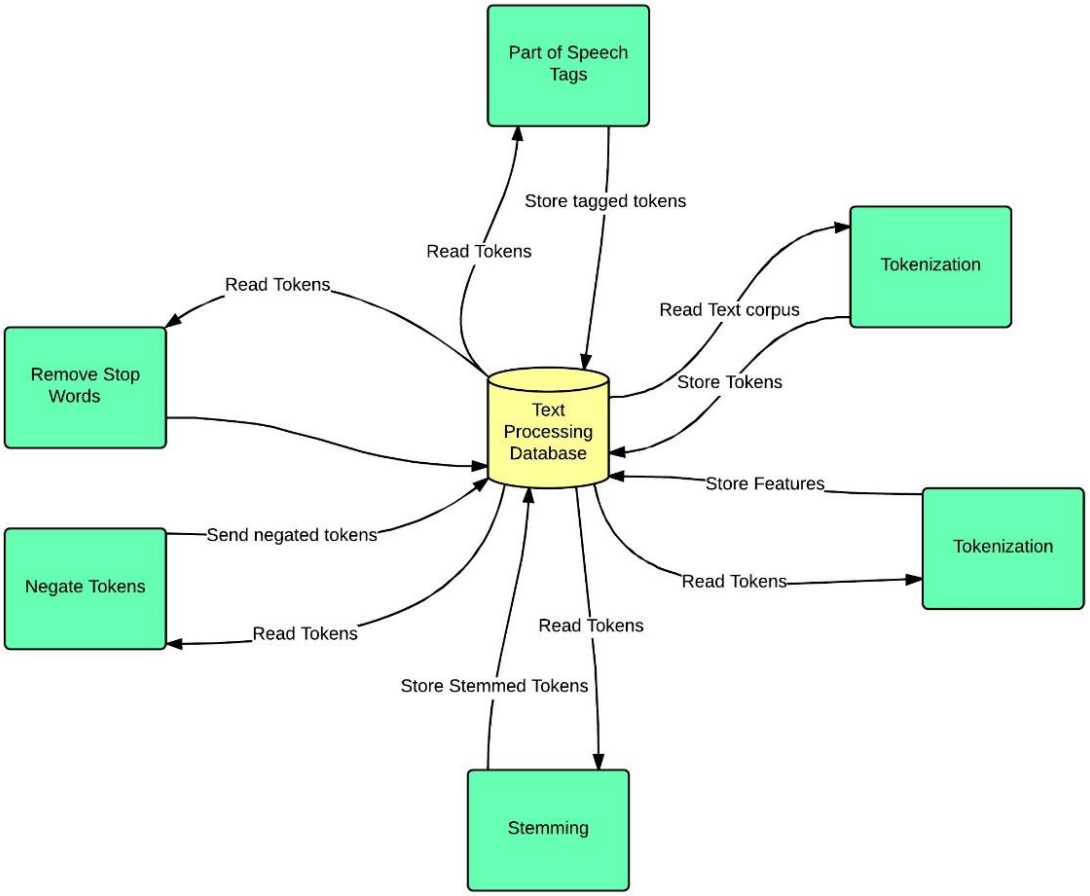


Figure 2.3 Text Pre-processing

- Tokenization: tokenization breaks a text unit (document, sentence) into individual words (tokens), the output is a tokens text corpus (Nicholls & Song, 2010; Zhang, 2013).
- Remove stop words: this step removes stop words like “a”, “the”, “are”, URLs, #, @ from the tokens text corpus. Removing these words, will reduce the size of the tokens text corpus, and therefore leads to less noisy tokens (tokens which don’t add meaning to the text corpus) (Nicholls & Song, 2010; Zhang, 2013).

- Text stemming: this step is applied in many research studies as for example in (Azar 2009; Siering 2012a). This technique stems tokens to their roots, to reduce the size of the tokens text corpus (Loughran and McDonald, 2011). Examples of online text stemmers are Porter Stemmer (Gupta & Lehal, 2013), Lancaster Stemmer (Zamora, 2016), WordNet Stemmer (Esuli & Sebastiani, 2006).

- Negation: common expressions like “not”, “not only”, “but” negates the meaning of the token/word preceding it. Example “I don’t like to drive Toyota cars”, this sentence is expressing negative opinion about his/her experience driving Toyota. Including this step improves the accuracy of understanding the text corpus (Das and Chen, 2007).

- Part of Speech Tagging (POS): this step is used normally when analyzing a text corpus using specific techniques such as natural language processing techniques (Kouloumpis et al., 2011). POS finds and tags each sentence’s nouns, verbs, adverbs, prepositions and adjectives. Some of the popular POS tagged corpuses is Trigrams'n'Tags (TnT) (Brants, 2000) and Brown Corpus. prepared by Brown University. (Atwell, 2016).

- Features extraction: this step reduces the text corpus in size by cutting down the features (words/tokens) that are not important. The importance of a token is decided using statistical techniques, where each token is assigned a statistical weight, and weighted against other tokens in the corpus. A variety of statistical based sentiment classification techniques (Medhat et al., 2014) utilize the top *n* features found in the corpus to calculate the text corpus sentiment scores. Some of these features extraction techniques are: unigram/bigram features, term frequency, inverse term presence, inverse document frequency, term frequency–inverse document frequency (Pang et al., 2002; Manning et al., 2008), mutual information (Sebastiani et al., 2002), information gain (Yang & Pedersen, 1997) and X^2 Feature selection (Rice, 2006). Instead of extracting features from text using these techniques, some studies use lexicons (dictionaries with prior polarity of positive, neutral and negative), or consider the tagged tokens produced by previous step (Part of speech) as the set of important features (Kouloumpis et al., 2011). Table 2.1 below surveys sample of research studies where pre-processing techniques were implemented.

Table 2.1 Pre-processing step used in a sample of sentiment analysis studies

	Tokenization	Remove Stop Words	Stemming	Features Extraction	Using External Lexicon based features
(Yang Yu et al., 2013)	✓			✓	
(Siering, 2012)		✓	✓	✓	
(Schumaker et al., 2012)		✓		✓	
(Azar, 2009)	✓	✓	✓	✓	
(Loughran and McDonald, 2011)	✓				✓
(Kothari et al., 2009)			✓		✓
(Tetlock et al., 2008)					✓
(Tetlock, 2007)					✓
(Das and Chen, 2007)			✓	✓	✓
(Antweiler and Frank, 2004)				✓	
(Mittermayer, 2004)	✓		✓	✓	
(Zhang, 2013)	✓	✓		✓	
(Hagenau et al., 2013)	✓	✓	✓	✓	

2.2.3 Calculate Sentiment Metrics (CSM)

This process is designed to read a text corpus (with features already extracted) and analyse the sentiment orientation of each element in the corpus. This process will generate sentiment metrics consisting of sentiment weight and a sentiment class (positive, negative, neutral) for each element in

the text corpus. Four main approaches are found in literature, these are: knowledge based, machine learning, natural language processing and semantic processing.

2.2.3.1 Knowledge based approaches

Knowledge based approaches rely on some knowledge base to determine sentiment. We distinguish two types of knowledge bases: dictionaries and ontologies. These approaches define the set of features (highly important affective words) to generate the sentiment metrics. Dictionaries are a collection of words/features with their sentiment orientation (negative, positive or neutral) such as General Inquirer lexicon (General Inquirer, 2016), finance sentiment words dictionary authored by (Loughran and McDonald, 2011) and Henry's finance-specific dictionary (Henry, 2008). Numerous examples from literature have utilized such dictionaries, as in the works of (Tetlock, 2007; Tetlock et al., 2008; Loughran & McDonald, 2011; Feuerriegel & Neumann, 2014). Another knowledge base can be built using ontologies, which are concerned with the semantic part of analysing a text corpus. These ontologies often contain huge resources of affective concepts and their semantics where each word and its synonyms, hyponyms, sentiment orientation, and sentiment weight is gathered. For example, Figure 2.4 shows an example of concept *"celebrating special occasion"* and its related semantics and polarity (Cambria et al., 2013/2014). Examples of sentiment based ontologies are: SenticNet (2014), Microsoft's Probose, Princeton's WordNet and MIT's ConceptNet (Cambria et al., 2012). Both dictionaries and ontologies can be used to generate sentiment metrics using these knowledge bases to assign a sentiment score (weight) and sentiment class (positive, negative) for each document in the corpus (e.g. Tetlock et al. (2008); Chatzakou et al. (2015)).

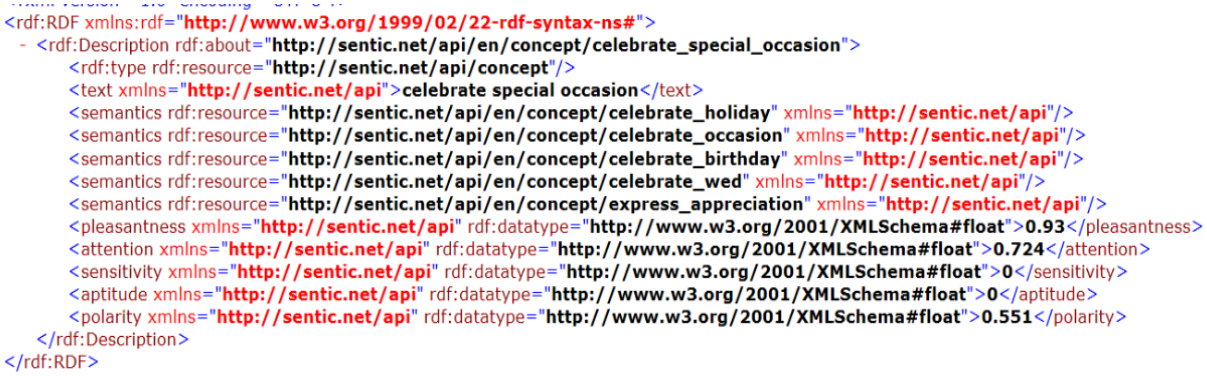


Figure 2.4 Example of SenticNet using ontology based approach (SenticNet, 2014)

2.2.3.2 Machine Learning (ML) approaches

Machine Learning (ML) approaches utilise algorithms that are able to perform automated text analysis including sentiment analysis of a text corpus. ML approaches rely on a pre-classified features matrix (tokens with their statistical weights) which can be generated using the text pre-processing steps (see subsection 2.2.2). The pre-classified corpus is often called a training set, which is a subset of the text corpus (the whole entire set of documents). To analyse sentiment, ML take into consideration the training set and the features matrix to classify the remaining (unclassified) text corpus (often known as the test set). Some of the popular approaches used to perform sentiment analysis are: Naïve Bayes (Medhat et al., 2014; NLTK, 2016), Decision trees (Azar, 2009), Artificial Neural Networks (Gershenson, 2003; Medhat et al., 2014), Support Vector Machines (Siering (2012a/2012b) and Medhat et al. (2014)).

2.2.3.3 Natural Language Processing (NLP) approaches

ML approaches apply statistical models to classify text document/corpus, which ignore the linguistic rules found in text. In contrast, NLP approaches take into consideration linguistic rules to make sense of the meaning. These rules include Part of Speech (POS) tagging, co-referencing, morphological analysis, stemming, parsing, and name phrase chunking. (Wu et al., 2009). For example, (Jakob and Gurevych, 2010) used POS tagged corpus and tokenization to extract aspects or sentences that carry sentiment from single and cross domain customer reviews. In addition, these NLP approaches have been used in a number of common NLP tools such as: Apache Open NLP (2016), Apache UIMA (2016), Stanford NLP (2016), GATE (2016), Mallet (2016), Natural Language Toolkit (NLTK, 2016), and Scala NLP (2016).

2.2.4 News analytics datasets

Some providers make use of their large text corpuses by generating news analytics datasets using proprietary technology. For example, Thomson Reuters provides a service since 2010, called Thomson Reuters News Analytics (TRNA), where news sentiment for each of the Reuters news story is calculated and provided to their clients. These sentiment datasets can be used in studying the impact on financial markets (Thomson Reuters, 2014; Rahman, 2014). There are other similar products by companies such as RavenPack News Analytics (2016), Quandl (2016) and Bloomberg Sentiment Data (2016).

2.2.5 Sentiment analysis accuracy metrics

Sentiment datasets accuracy can be determined by various metrics such: precision, recall and accuracy (Inkpen, & Désilet (2005); Siering (2012a/2012b)). They use a subset of the main corpus which researchers select and manually classify (Azar, 2009). This subset is then used in evaluating the performance of the classification accuracy of the remaining set.

2.2.5.1 Precision

High precision figure means less false positives (StreamHacker, 2010). The list of false positives is attained by comparing the sentiment classifier decision (list of positives) with the hand annotated set. Precision figure is reached using the following equation.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2.1)$$

Where *TP* is the number of true positives samples, *FP* is the number of false positives samples.

2.2.5.2 Recall

Recall is very important measure; it measures the sensitivity of classifiers to miss classifying documents (StreamHacker, 2010). Recall is associated with the number of false negatives. False negatives are the number of falsely assigned negative class, and it is calculated using the following equation.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2.2)$$

Where *TP* is the number of true positives samples, *FN* is the number of false negatives.

2.2.5.3 Accuracy

Accuracy measures the accuracy percentage using the following equation.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (2.3)$$

Where *TP* is the number of true positives samples, *TN* is the number of true negative. *FP* is the number false positives and *FN* is the number of false negative samples.

2.2.5.4 F1 Score

F1 score measures the accuracy of a test by combining the precision and recall together to compute a sentiment score (StreamHacker, 2010). It can be calculated using the following equation.

$$\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (2.4)$$

2.3 Overview of financial markets

This section covers a brief overview of the financial markets, definitions, theories, types of market data, and their measures.

2.3.1 Financial markets definitions

Financial markets are marketplaces where financial instruments such as equities, bonds, currencies and derivatives are traded (Viney ,2003). Stock markets (equity or share markets) are a type of financial markets where traders can sell/buy shares of publicly listed companies. For example, Google, IBM or Microsoft shares can be traded on New York Stock Exchange or London Stock Exchange. Financial markets group stocks or bonds to form market indices. Groupings could be on the basis of capital size (large, medium or small) or industry sector (technology sector, tourism, health …etc.) (InvestingAnswers, 2017b). At any time, an index has a value, which is determined by calculating the average price/market capitalization of the companies included on the index. These index values provide investors an idea on the performance of the group of stocks that belong to the index. A popular example of a market index is DJIA (Dow Jones Industrial Average), which is a market index listing the top thirty (by company size) companies in the US, such as Microsoft, HP, IBM. Other examples of popular indices All Ordinaries Index listed on Australian Stock Exchange, German DAX Index listed on Frankfurt Stock Exchange. Market indices.

The huge influx in information available through various heterogeneous resources (newswires, blogs, messages boards, social media) makes it complex for traders (fund managers, investors) to explain stock price movements. Researchers have investigated the relationship between the financial markets and news for around five decades (Niederhoffer, 1971; Antweiler and Frank, 2004). There are many theories that have been proposed trying to explain this relationship including Efficient Market Hypothesis, Behavioural Economics, and Adaptive Market Hypotheses.

Efficient Market Hypothesis (EMH) assumes that markets are very efficient in responding to news and therefore, the prices within the markets reflect the news almost instantly (Choi, 2013). EMH has three different forms, strong form, semi-strong form and weak form. The strong form supports the argument that financial markets are 100% efficient and therefore the room to make money by capitalizing on information is almost zero, because all information is public and reflects the current value of the stock. The semi strong form suggests that making money is only possible if the investor researches the company thoroughly before trading. The weak form says that the markets are in fact inefficient and they don't reflect all the information available and scenarios of insider trading are proof of this situation.

Behavioural Economics theories investigate the cognitive parts of the decision-making process. They focus on the drivers behind making buying or selling decisions in the market place. These theories challenge the efficient market hypothesis by arguing that markets are driven by human emotions such as fear and greed. Behavioural theories investigate motives, reactions, overconfidence, overreaction and underreaction patterns and try to explain them. The theories firmly established that investors behaviour is shaped by how optimistic or pessimistic they feel about the future value of their stock/market of interest (Bollen and Mao, 2011).

Adaptive Market Hypothesis marry efficient market hypotheses with the behavioural/cognitive theories (Lo, 2004). These theories assume that investors usually adapt their behaviour to the situation depending on the surrounding circumstances, where they base their actions on trial and error. For example, if one investment strategy fails, they adapt and try another one until they find a successful strategy and this is likely to be repeated again.

2.3.2 Financial markets data

Financial markets generate a large amount of market data, which several major data providers (such asThomson Reuters, Bloomberg and Quandl) collect, aggregate, and distribute to their clients. Market data can be of different frequencies e.g. low, medium or high frequency. Monthly and weekly data is considered low frequency, whereas daily data is considered medium frequency and high frequency refers to market data timestamped on real-time basis. Most common market data types are:

- End of day shows the daily opening, closing, highest and lowest prices for a particular stock. Investors use this type of data to evaluate the performance of the stocks, comparing the opening and closing prices to understand the stocks trends (Just Data, 2017).
- Quotes shows the best bids and asks and their size offered by market makers (brokers, traders) for a particular stock (InvestingAnswers 2017a).
- Trades are transactions of exchanging shares or bonds between the market makers. Trades are usually of high frequency (stored with the actual time of the transaction down to the millisecond). Trades and quotes are kept in an order book generated by stock exchanges around the world. These order books keep track of trades, and quotes of bids (buyers side) and asks (sellers side) (Milton, 2016).
- Index values represent the average value of multiple stocks' prices calculated for a group of companies which are listed on an index.
- Market Depth data shows different levels of quotes related to the same stock (Milton, 2016).

2.3.3 Market measures

The financial markets are characterized by a set of measures that reflect the trading activity conditions. This section briefly explains these the most known indicators (Baker and Wurgler, 2006).

2.3.3.1 Stock Returns

Stock returns are calculated to provide an indication of the value of a particular stock (French, 1980; Brown and Warner, 1985; Tetlock, 2007). Stock returns can be calculated using daily trading prices or higher frequency trading prices such as seconds, minutes, or hours (e.g. Schumaker et al., 2012; Lugmayr and Gossen, 2013). The equation is illustrated as follows:

$$R_t = \frac{(\mathrm{Pt} - \mathrm{Pt}-1)}{(\mathrm{Pt}-1)} \qquad (2.5)$$

Where R_t is the returns of time t. P_t the trading price of a stock at time t. P_{t-1} the trading price of a stock at time $t-1$.

2.3.3.2 Volatility

Volatility is a statistical measure of the dispersion of a stock compared to a benchmark. It is usually often used to measure the risk or uncertainty associated with a particular stock. This dispersion can be measured using the standard deviation of stock returns (of a particular stock) from the mean of stock returns (usually a portfolio or a market index). The higher the dispersion is, the more volatile this stock will be. Other popular models have been used widely in the financial industry and in the applied econometrics literature to calculate volatility of stock returns, these are ARCH and GARCH models, which stand for autoregressive conditional heteroskedasticity and generalized autoregressive conditional heteroskedasticity (Engle, 2001). Stock returns of one, two or three deviations of the mean are deemed noteworthy, as they could indicate events or news that released on those days, have caused the stock to experience high volatility.

2.3.3.3 Trading volume and liquidity

Trading volume provides information about the number of shares/contracts traded during a time period (for example: weekly, daily or intraday basis). A volume spike is when the volume is doubled or tripled of that the previous days (Rockefeller, 2016). Trading volume is a measure strongly correlated with news (Das and Chen, 2007), and it has been used as a direct evidence of increased/decreased demand on the stock. Liquidity is a measure strongly connected to trading volume. It is defined as the level in which an asset can be sold or bought in the stock market. High liquidity (also known liquid assets) means the stock has high demand and low liquidity indicates weak demand (Investopedia, 2016b). Liquidity can be calculated using the following measurement ratio:

$$Current\ Ratio\ = \frac{Current\ Assets}{\text{Cur}rent\ Liabilities} \qquad (2.6)$$

2.4 Sentiment analysis in finance research

Researchers tapping into finance and computing domains need a good understanding of the topics covered in section 2.2 related to the techniques employed to generate sentiment datasets. In addition, they require to understand information related to financial markets, existing data types and measures as briefed in section 2.3. Therefore, the thesis extends the literature review by presenting example studies that attempt to explain the relationship between sentiment found in text (e.g. news articles, tweets…etc.) and the performance of the financial markets. In this section, we review first the different

sentiment analysis techniques used in finance research studies and then review the financial impact analysis techniques that they have used.

2.4.1 Sentiment analysis techniques used in finance studies

In subsection 2.2.3 the thesis covered three main approaches found in sentiment analysis literature, these were the knowledge based, machine learning, and natural language processing approaches. In this subsection, we overview the studies that utilised these approaches in the finance domain.

2.4.1.1 Measuring sentiment using knowledge based approaches

In studies using a knowledge base, most of them utilized a dictionary or a statistical text analysis tool that utilized dictionaries. Of these some utilized the Harvard General Inquirer text analysis tool (General Inquirer, 2016), which is based on Harvard dictionary (e.g. Engelberg (2008); Feldman et al. (2008)). Others utilized DICTION text analysis tool (Digitext, 2017) (e.g. Davis et al. (2012); Davis and Tama-Sweet (2012); Demers and Vega (2014)). Other studies relied on word lists specifically authored for finance research (e.g. Henry (2008); Henry and Leone (2009); Loughran and McDonald (2011); Doran et al. (2012); Engelberg et al. (2012); Huang et al. (2013); Davis et al. (2015)). The most common sentiment measure is to compare the number of words in a given sentiment class (positive or negative) with the total number of words found in the text corpus (e.g. Kothari et al. (2009); Chen et al. (2013); Ferguson et al. (2015)). Tetlock et al. (2008) disagreed with this approach and used the General Inquirer (Harvard) tool to assign a sentiment score for each of the news articles they analysed. They claimed that the positivity of a news message is better represented as messages with less negativity ratio. This is because counting positive words is not as accurate as many of the positive words are used with negations. i.e. words like “not good” could be counted as positive but the meaning of the phrase is in itself negative. Using negative words with negations is less frequent as (Tetlock, 2007) found. Loughran and McDonald (2011) went a step ahead and calculated a weight for each financial term based on corporate filing reports extracted from EDGAR (2017) website. Their approach weighed for example a word like “buy” as heavy as a word like “position”, because the number of occurrences of the word “buy” was more than the number of occurrences of the word “position”. These weights formed the basis for calculating a sentiment score for each filing report.

2.4.1.2 Measuring sentiment using machine learning approaches

Computer scientists applied more sophisticated techniques to construct sentiment datasets utilising machine learning approaches (e.g. Antweiler and Frank (2004); Das and Chen (2007); Li (2010); Siering (2012a); Huang et al. (2013)) which is based on statistical learning theories to analyse the contents of a text corpus. The general methodology followed was to designate a portion of the text corpus as a training set, and the remainder as a test set. Each word in the training set is manually classified (positive, negative, or neutral) (Li, 2010) or classified using unsupervised technique such as in (Siering, 2012a), which relied on a dictionary to classify the words in the training corpus. Some of the popular works utilising machine learning approaches like (Das and Chen, 2007) applied a voting scheme between five different sentiment classification algorithms, and selected the best sentiment scores relying on statistical confidence test as a measure of most and least accurate classifiers.

2.4.1.3 Summary of datasets used

The reviewed studies utilized two essential data types which are: market data and text data. Market data varied between daily and intraday frequencies. Text data varied in their types, some used scheduled and others utilized unscheduled sources. Some researchers have used social media sources, such as (Antweiler and Frank 2004; Das and Chen, 2007) where they incorporated blog messages and others such as in (Bollen and Mao, 2011; Vu et al., 2012) relied on text from Twitter to predict markets' movements. Another group used less noisy sources (Huang et al., 2010) where they utilised scheduled news announcements e.g. Henry (2008); Feldman et al. (2008); Henry and Leone (2009); Li (2010); Loughran and McDonald (2011); Davis and Tama-Sweet (2012); Doran et al. (2012); Price et al. (2012); Huang et al. (2013); Davis et al. (2015). The studies that utilised unscheduled news sources such as blog messages, tweets are summarised in Table 2.2. In Table 2.3 studies which used scheduled news sources such as earning announcements, corporate disclosures, or 10k filing reports are shown.

Table 2.2 Unscheduled news sources

Research Study		Text Data		Market Data		
		Text Type	Text Source	Time Frame	Sample data period length	Financial Markets Data Sources
1	(Antweiler and Frank, 2004)	Blog message	Yahoo Finance and Raging Bull	Intraday	1 year	45 companies listed on Dow Jones Industrial Average (DJIA) index
2	(Das and Chen, 2007)	Blog message	Messages from stock boards	Daily	2 months	24 tech-sector stocks, listed on the Morgan Stanley High-Tech index (MSH)
3	(Tetlock, 2007; Tetlock et al., 2008)	Stock news articles	Dow Jones Wall Street Journal	Daily	24 years	S&P 500 listed companies stock data
4	(Engelberg, 2008)	News articles	Dow Jones News Service (DJNS)	Daily	7 years	4700 US companies from Center for Research in Securities Prices (CRSP)
5	(Bollen and Mao, 2011)	Tweets	Twitter	Daily	10 months	DJIA index
6	(Dzielinski, 2011)	Stock news articles	Thomson Reuters	Daily	7.5 Years	950 companies trading on NYSE and NASDAQ
7	(Schumaker et al., 2012)	Stock news articles	Yahoo finance	Intraday	1 month	S&P 500 listed companies
8	(Vu et al., 2012)	Tweets	Twitter	Daily	2 months	4 tech companies NASDAQ stocks
9	(Engelberg et al., 2012)	News articles	Dow Jones archive	Daily	2.5 years	3167 NYSE-listed companies
10	(Siering, 2012a)	Stock news articles	Dow Jones News	Intraday	1 year	German DAX 30 Index
11	(Siering, 2012b)	Stock news articles	Dow Jones News	Daily	14 months	Dow Jones Industrial Average (DJIA) index

Research Study		Text Data		Market Data		
		Text Type	Text Source	Time Frame	Sample data period length	Financial Markets Data Sources
12	(Allen et al., 2015)	Stock news articles	Thomson Reuters	Daily	6 Years and 10 months	30 companies listed on DJIA index

Table 2.3 Scheduled news sources

Research Study		Text Data		Market Data		
		Text Type	Text Source	Time Frame	Sample data period length	Financial Markets Data Sources
1	(Mittermayer, 2004)	Corporate announcements	PR Newswire	Intraday	1 year	Companies trading on NYSE, NASDAQ, AMEX, and 5 regional stock exchanges with turnover of at least US $ 5,000,000 a day (averaged over year 2002)
2	(Feldman et al., 2008)	10K filing reports	Compustat database	Daily	11 years	All companies trading on NYSE, AMEX or NASDAQ exchanges during study period
3	(Henry, 2008)	Corporate announcements (Earnings)	Lexis-Nexis or Factiva websites	Daily	5 years	562 companies found in CRSP database
4	(Henry and Leone, 2009)	Corporate announcements (Earnings)	EDGAR website	Daily	3 years	562 companies found in CRSP database
5	(Kothari et al., 2009)	Disclosure reports	Corporates and Analysts report	Daily	6 years	Dow Jones and Factiva
6	(Davis et al., 2012)	Corporate announcements (Earnings)	PR Newswire	Daily	6 years	542 Companies in CRSP and Compustat databases that are mentioned in the earnings announcements

7	(Loughran and McDonald, 2011)	Corporate Filing reports	EDGAR website	Daily	14 years	8341 companies listed on NYSE, NASDAQ, or AMEX
8	(Davis and Tama-Sweet, 2012)	Corporate announcements (Earnings) and 10K filings	PR Newswire and Morningstar Document Research	Daily	6 years	542 Companies in CRSP and Compustat databases
9	(Doran et al., 2012)	Corporate announcements (Earnings)	Fair Disclosure Wire and The American Intelligence Wire	Daily	4 years	233 Stocks listed on the National Association of Real Estate Investment Trusts in USA
10	(Price et al., 2012)	Corporate announcements (Earnings)	Thomson Reuters' First Call Historical database	Daily	4 years	Companies related to 2880 earnings announcements derived from CRSP-COMPUSTAT Merged database
11	(Jegadeesh and Wu, 2013)	10K filings	EDGAR website	Daily	6 years	7606 companies found with 10k filings in CRSP-COMPUSTAT Merged database
12	(Hagenau et al., 2013)	Corporate announcements	Deutsche Gesellschaft für Adhoc-Publizität (DGAP), EuroAdhoc websites	Daily	14 years	Selected sample of companies listed in Germany and UK stock exchanges
13	(Demers and Vega, 2014)	Corporate announcements (Earnings)	PR Newswire	Daily	9 years	2729 companies with earnings announcements in Compustat database
14	(Feuerriegel & Neumann, 2014)	Corporate announcements	Deutsche Gesellschaft für Adhoc-Publizität (DGAP)	Daily	7.5 years	485 companies listed on CDAX index in Germany

15	(Davis et al., 2015)	Corporate announcements (Earnings)	CQ FD Disclosure database available on Factiva	Daily	8 years	225 US companies found in Compustat database

Market data collected varied between intraday (5 min to 15 min to 1 hour) intervals (e.g. Antweiler and Frank, 2004; Mittermayer, 2004; Sieirng, 2012; Schumaker et al., 2012) to daily market data. Daily based market data is the most common frequency among the studies reviewed. Knowledge based approaches were the most common in studies that used scheduled news sources, while machine learning approaches were more popular among studies that used unscheduled news sources.

2.4.2 Techniques used to evaluate impact

The studies reviewed have followed one or a combination of two main methodologies to evaluate the impact of the sentiment data on the market data. These are: regression analysis (e.g. Linear Regression (LnR), Ordinary Least Squares (OLS) regression, Logistic Regression (LoR)) and trading strategies (see Table 2.4).

Reviewed studies show that researchers with strong finance background often prefer to use regression analysis models. Linear regression was the most popular impact analysis methodology in the reviewed studies. These studies often define the independent variable as the event they are investigating (e.g. earning announcement, analyst report, news article), and the dependent variable to be a market measure such as future earning, stock returns, trading volume etc (e.g. Engelberg (2008); Feldman et al. (2008); Dzielinski (2011); Loughran and McDonald (2011); Engelberg et al. (2012); Price et al. (2012)). Linear regression is often used in conjunction with an event study methodology to adjust the evaluation results (Kothari and Warner (2004); Corrado (2011)).

Other researchers (e.g. Tetlock, 2007; Tetlock et al.,2008; Allen et al., 2015) applied more advanced regression models such as Ordinary Least Squares (OLS) regression. This method can analyze the interdependencies of many variables. It has been used to understand interdependencies between company performance indicators (such as stock returns), sentiment data and other control variables (trading volume, index returns, insider trading). In other studies (e.g. Loughran and McDonald (2011)), logistic regression can use sentiment data extracted from analyzing corporate reports to predict if certain events are going to occur (corporate fraud, weakness disclosures).

Trading strategies is another methodology considered as a complementary technique to other regression analysis techniques (e.g. Tetlock et al. (2008); Loughran and McDonald (2011); Schumaker et al. (2012); Engelberg et al. (2012); Siering (2012a/2012b); Hagenau et al. (2013); Demers and Vega (2014)). For example, Engelberg (2008) and Tetlock et al. (2008) both used a simple trading strategy to buy and hold on positive news and sell on negative news. Abnormal returns were examined based on trading strategies constructed by other researchers (e.g. Loughran and McDonald (2011), Engelberg et al. (2012); Siering (2012a/2012b)). Others predicted the trading signals direction (Schumaker et al., 2012).

In general, the reviewed studies have evaluated the impact of their sentiment data using one or a combination of market measures. Stock prices and stock returns were the most studied market measure (e.g. Engelberg (2008); Feldman et al. (2008); Dzielinski (2011); Loughran and McDonald (2011); Engelberg et al. (2012); Price et al. (2012); Allen et al. (2015)). A volatility measure was used in some studies (e.g. Antweiler and Frank (2004); Das and Chen (2007); Kothari et al. (2009); Loughran and McDonald (2011)). Other studies used a liquidity measure such as trading volume (e.g. Antweiler and Frank (2004); Das and Chen (2007); Tetlock (2007); Tetlock et al. (2008); Siering (2012b); Jegadeesh and Wu (2013)).

Table 2.4 Sentiment analysis approaches and impact models

Research Study		**Sentiment analysis approach**			**Impact model**				**Findings**
		Balanced Data	**Machine learning**	**Knowledge based**	**Regression analysis**				
					LnR	LoR	OLS	**Trading strategy**	
1	(Antweiler and Frank, 2004)	No	✓		✓				A positive shock to message board postings predicts negative returns the next day; contemporaneous regressions show that disagreement induces trading; message posting helps predict volatility; stock messages reflect

									public information very rapidly.
2	(Mittermayer, 2004)	Yes	✓					✓	Average profit 11% compared to average profit by random trader 0%
3	(Das and Chen, 2007)	No	✓		✓				Sentiment aggregated across stocks' tracks index returns; aggregation of sentiment reduces some of the noise from individual stock postings; market activity is related to small investor sentiment and message board activity.
4	(Tetlock, 2007) and (Tetlock et al., 2008)	Not mentioned		✓	✓			✓	Negative words convey negative information about firm earnings above and beyond stock analysts' forecasts and historical accounting data; stock market prices respond to the information embedded in negative words with a small, one-day delay; negative words in stories about fundamentals predict earnings and returns more effectively than negative words in other stories.
5	(Engelberg, 2008)	Not mentioned		✓	✓			✓	Qualitative earnings information embedded in the news stories about firms' earnings announcement has additional predictability for asset prices beyond the quantitative

									information; qualitative information about positive fundamentals and future performance.
6	(Feldman et al., 2008)	Not mentioned		✓	✓			✓	Changes in the tone of the Management Discussion and Analysis section from the recent past are significantly correlated with short window contemporaneous returns around the Securities and Exchange filing dates.
7	(Henry, 2008)	No		✓	✓				More positive tone in press releases results in higher Abnormal market returns
8	(Henry and Leone, 2009)	No		✓	✓				Domain-specific wordlist can significantly increase the performance of tests to measure the qualitative information in financial disclosure.
9	(Kothari et al., 2009)	Not mentioned		✓	✓				Business press impacts the markets more than analysts' forecasts. This is witnessed in the markets heavily discounting news disclosures from analysts ; while both positive and negative news disclosures in the business press shows higher impact on the cost of capital, return volatility, and analyst forecast dispersion

10	(Bollen and Mao, 2011)	Not mentioned	✓		✓				The results strongly indicate a predictive correlation between Twitter mood and DJIA values, they offer no information on the causative mechanisms.
11	(Dzielinski, 2011)	No	✓				✓		News days (positive/negative) indicate higher/lower returns than no-news days.
12	(Loughran and McDonald, 2011)	Not mentioned		✓	✓	✓		✓	Results show words that appear on Barron's list significantly related to excess filing period returns, analyst earnings forecast dispersion, subsequent return volatility, and fraud allegations
13	(Davis et al., 2012)	Not mentioned		✓	✓				Managers' earnings press release language communicates credible information about expected future firm performance to the market and that the market responds to this information
14	(Davis and Tama-Sweet, 2012)	Not mentioned		✓	✓				Higher levels of pessimistic language in the Management Discussion and Analysis are associated with lower future return on assets, which managers try to omit in their earnings announcements
15	(Doran et al., 2012)	Not mentioned		✓	✓				Using the customized dictionary as in Henry (2008) and Loughran and McDonald, (2011)

									provide significantly strong explanatory power for the accompanying abnormal returns.
16	(Engelberg et al., 2012)	Not mentioned		✓	✓			✓	Strategy based on short selling and negative news would have earned an astonishing 180% during the authors' 2.5-year sample period
17	(Price et al., 2012)	No		✓	✓				Conference call discussion tone has highly significant explanatory power for initial reaction window abnormal returns as well as the post-earnings-announcement drift
18	(Schumaker et al., 2012)	No	✓					✓	Trading strategy based on subjective articles performed good with 59.0% directional accuracy and a 3.30% trading returns as compared to poor performance in the case of objective articles.
19	(Siering, 2012a)	Yes	✓		✓			✓	Dictionary authored by Loughran and McDonald, (2011) performed provided better statistically significant stock returns than Harvard General Inquirer dictionary,
20	(Siering, 2012b)	Yes	✓		✓			✓	Increase in Investor attention lead to better returns than relying only on sentiment orientation of news.

21	(Vu et al., 2012)	Yes	✓				✓		Strong reaction to the features selected from tweets related to four large tech companies, indicating high correlation between t tweets mood and the direction of the stock prices.
22	(Hagenau et al., 2013)	Yes	✓				✓	✓	2gram feature selection coupled with Feedback-based feature selection achieved accuracies of up to 76%
23	(Jegadeesh and Wu, 2013)	Not mentioned		✓	✓				Term weighing around filing reports dates found to be more correlated to returns than relying on sentiment word lexicons.
24	(Demers and Vega, 2014)	Not mentioned		✓	✓			✓	The authors found there is an inverse association between the certainty in management's diction and the idiosyncratic volatility in the company's share price during the announcement window.
25	(Feuerriegel & Neumann, 2014)	Not mentioned		✓				✓	Trading simulator significantly outperforms momentum trading strategy. and CDAX index
26	(Allen et al., 2015)	No	✓				✓		Significant relationship found between sentiment scores in TRNA dataset and DJIA constituents returns.

27	(Davis et al., 2015)	Not mentioned		✓	✓				Managers' specific characteristics (e.g., gender, age, educational and career experiences) potentially impact the market reaction to earnings announcements.

2.4.3 Discussion

The reviewed studies in subsections 2.4.1 and 2.4.2 show that researchers have utilised a variety of sentiment analysis techniques and impact models. These studies can be divided into one of the following two groups of research:

- First group of studies focused on proposing techniques which encapsulate processes to perform sentiment analysis for a specific text source, with limited impact analysis for validation purposes. The researchers in this group have advanced computing skills. Their research activities mainly focus around developing new sentiment analysis algorithms (i.e. machine learning). However, the corresponding impact measuring processes are often not documented and only accessible by the authors of the study. Users who wish to reuse part or all of their evaluation processes in different financial contexts, or for different news sources, hit a roadblock.
- Second group of studies have strong finance background, with domain expertise in financial markets modelling and evaluation. Their major focus is on evaluating the impact of a particular sentiment dataset using different financial market measures. Many of the evaluation activities are hard to reproduce for different sentiment datasets, as many of these activities are not automated. In addition, evaluation requires users to dedicate a great portion of their time to conduct these activities.

In conclusion, most of the results described in the literature concerning sentiment analysis perform impact analysis in a way that is difficult to reproduce outside of a specific context, i.e. the processes involved in evaluating the impact of a sentiment dataset are not documented. To reproduce a studies' results, one would need to implement the evaluation algorithms and have access to the input (news and market data) used. This makes the task of reproducing results a complex job (Jasny et al., 2011; Peng, 2011). The impact on financial markets can be gauged by several measures, but the majority of the

studies reviewed focused on studying the impact using one or two financial market measures. Stock returns (intraday, daily, monthly, annually) have been studied widely, and considered to be the most popular performance measure of impact on financial markets. However, the literature shows there are many other measures that could be indicators for impact such as liquidity, volatility (Lugmayr, 2013). The complexity of sentiment analysis models along with time limitation and/or limited knowledge in the financial markets domain, leads researchers to apply limited impact analysis evaluation.

The existing studies use a variety of statistical methods to test the impact sentiment had on financial market measures. Some use correlation tests, while others use regression analysis to discover the strength of the relationship between change in market measures and sentiment datasets. Others take one step further, implementing several regressions between the different financial market measures. In addition, most studies evaluated their impact models against one specific fixed financial context. However, this raises a question like *"what if the financial context or the financial market measure changes? would I get the same results?".* Automating the processes involved in analysing and evaluating data, whether it is market data, or sentiment data has many advantages: first reducing labour/manual work, reducing human errors, and reducing time and resources (Harcar, 2016).

2.5 Conclusion

This chapter, started by presenting an overview of the different processes and techniques involved into converting a text corpus into sentiment metrics/datasets. Then the chapter turned its attention to data needed as part of such processes like financial markets data types, events and measures. The chapter then reviewed some existing studies and assessed the techniques, datasets and impact models used by researchers. Lastly, this chapter identified the variety in conducting sentiment impact analysis and the lack of a systematic methodology for repeating a particular model over multiple contexts. In the next chapter, we will propose our research methodology, detailing the research questions, objectives and the methodology used in carrying out this research.

3 RESEARCH METHODOLOGY

In this chapter, we will first elaborate on the research problem in section 3.1, define the research questions in section 3.2, and then discuss the research approach in section 3.3. Next, the research process adopted in this thesis is described in section 3.4. Finally, section 3.5 concludes this chapter.

3.1 Research problem

The research gap discussed in subsection 2.4.3 had 3 different aspects:

- First aspect relates to the shortcomings of existing literature, where the lack of *flexibility* for conducting news sentiment datasets evaluations has been noticed. To understand the results in context, the task of deriving the parameters and datasets used in these studies is tedious and time-consuming. Often these parameters are either missing, or ambiguously defined. So, the first gap concerns the lack of *flexibility* in enabling users to derive and compare between different studies in terms of parameters, or datasets used.
- Second aspect relates to the absence of clear *step-by-step* guidance for replicating the results. Existing studies on sentiment driven impact analysis do not pay enough attention to this aspect. One reason being that the focus of many of these studies was not to produce a process-driven impact analysis study, but to produce results of high accuracy. User experience and usability aspects were not addressed, and there was little guidance on reproducing or reconducting a set of experiments. Therefore, there is no systematic methodology that has clear step-by-step use cases to enable *reproducibility* and *consistency* in conducting sentiment-driven impact analysis studies.
- Third aspect relates lack of software tools to support *automation* of such step-by-step use cases. The role of such tools would be to allow experiments to be conducted effectively and minimize the risk of errors.

3.2 Research questions

The research gap leads to the following research questions, which revolve around one main theme, that is investigating the feasibility of introducing a systematic method for conducting an impact analysis study. These questions are:

- "What are the parameters that uniquely define a context that enables validating sentiment datasets by conducting impact studies in multiple financial contexts?".
- "Given a context, what are the set of use cases that need to be defined to guide users to conduct their experiments in a consistent fashion?".
- "How to support automating these use cases with a single software framework?".

3.3 Research approach

To answer the research questions, we propose a software framework called News Sentiment Impact Analysis (NSIA), which comprises the following elements:

- A novel data model (Comparison Parameters Data) that captures contextual parameters in the financial markets and news sentiment analysis sphere. The data model simplifies the identification and representation of the parameters used in sentiment-driven impact analysis studies. This model will address the *flexibility* issue by providing users with a way to set different contexts for impact analysis
- A set of step-by-step predefined *use cases* to make the job of conducting experiments repeatable for the users. This will address the *reproducibility* issue by allowing users to repeat impact analysis studies.
- A *software architecture* that is designed to support both the data model, and the use cases associated with it. The software architecture should be able to facilitate *automation* of the impact analysis studies. The architecture also facilitates the reuse and *interoperability* of an existing software components, libraries, and packages in conducting impact analysis studies.

3.4 Research process

The research process follows an iterative cycle, which consists of 4 stages outlaid in Figure 3.1(Murch, 2001). The iterative cycle enables the delivered framework to continuously evolve and be reviewed against new requirements that emerge during analysis, design, implementation, or evaluation stages (Anderson, 2000). The four stages are illustrated as follows:

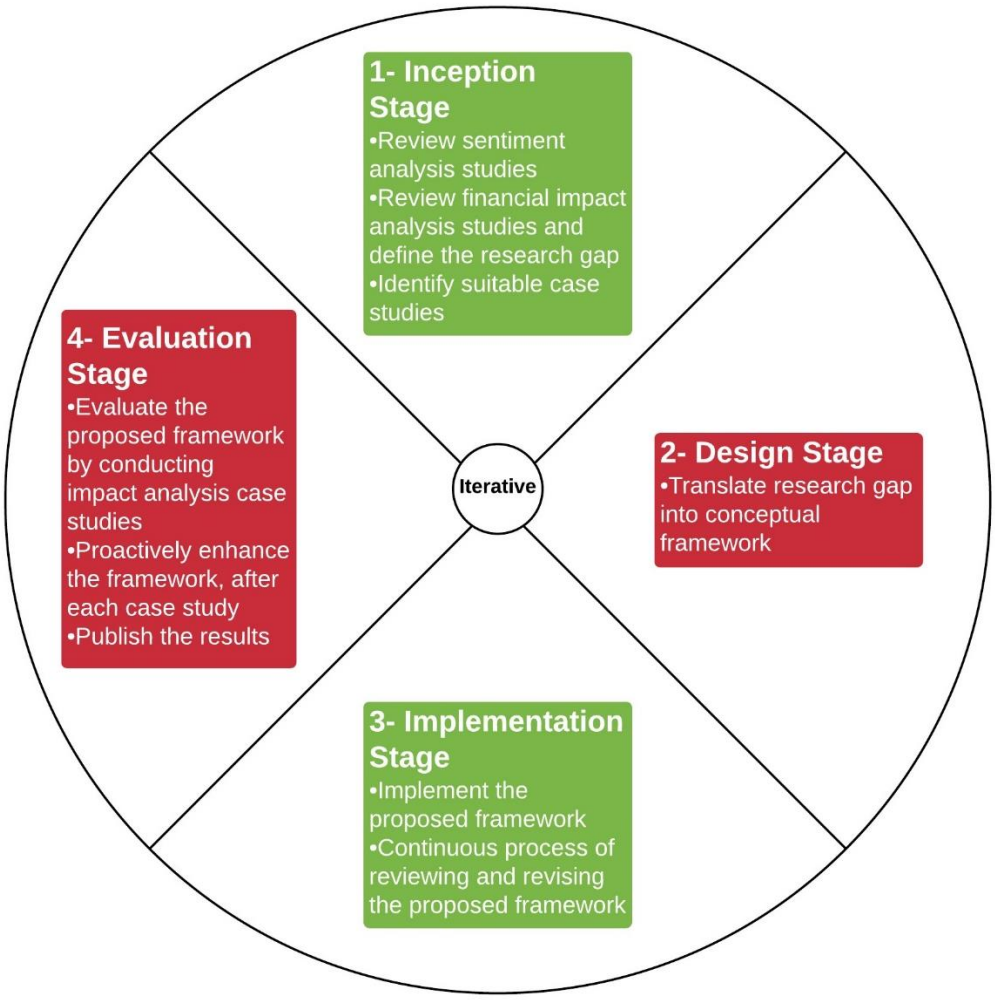


Figure 3.1 Research Stages in research process

- Inception stage: In this phase, a literature review is prepared, investigating existing research efforts around sentiment analysis methods and sentiment analysis applications to finance. The outcomes of this phase are *well-defined* requirements that are not adequately addressed by existing techniques. The findings of this stage have been presented in chapter 2.
- Design stage: This research activity aims at translating the outcomes of the inception stage into design artefacts. The outcome of this stage is the News Sentiment Impact Analysis (NSIA) framework, which will be described in chapter 4.
- Implementation stage: A prototype is developed to test the effectiveness of NSIA framework. The prototype aims at validating the functionality and feasibility of the proposed framework. More details on the implementation will be given in chapter 5.

- Evaluation stage: The research process will be validated using case studies in chapter 6.

Choosing case studies as the methodology to validate our research process is motivated by the following reasons:

- Research on sentiment driven impact analysis is a "mixed method" activity, as it implies both qualitative and quantitative data (Runeson and Höst, 2009). Where qualitative information (news, blogs) are translated into sentiment indicators. This makes the utilization of case studies methodology a suitable methodology to evaluate the effectiveness of the framework.
- Case studies are suitable methodology to explain the pre- and post- events used in *time-series* analysis (Runeson and Höst, 2009). The work conducted in this thesis utilizes *time-series* analysis as technique to analyse the events overtime, for the purpose of understanding the effect sentiment data had on the performance of a financial entity.
- The case studies' scenarios in this thesis work utilize real data from the financial markets, and financial news domain. These studies will help us determine if the proposed framework is in fact applicable in the real world (Runeson and Höst, 2009; Shuttleworth, 2016).

3.5 Conclusion

This chapter summarized the research problem, presented the research questions, research approach and research process used in this thesis. A framework called News Sentiment Impact Analysis (NSIA) is proposed. The research process will centre around the design implementation and testing of the framework. The next chapter will describe the proposed framework and its components in more detail.

4NSIA FRAMEWORK

This chapter introduces the News Sentiment Impact Analysis (NSIA) framework which corresponds to the design stage stated in the research process. First, section 4.1 gives an overview of the NSIA framework. Section 4.2. describes the first component which is a novel conceptual data model that addresses the lack of flexibility in existing sentiment-driven impact analysis studies. Section 4.3 discusses the second component which is a software architecture to enable delivery of the data model. Both the data model and the software architecture are managed by a set of well-defined use cases detailed in section 4.4. Lastly, section 4.5 concludes this chapter.

4.1 Overview of the NSIA framework

The NSIA framework is composed of three main components, each component corresponds to one of the design artefacts defined in section 3.3. The framework comprised of a novel data model, a software architecture, and a set of use cases as shown in Figure 4.1.The rest of this section describes these components in more detail.

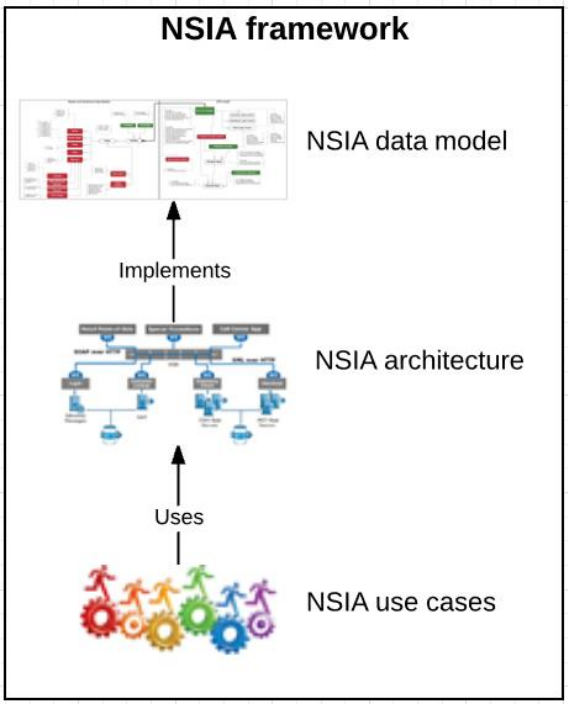


Figure 4.1 NSIA framework components

4.2 NSIA data model

The NSIA data model is made up of three distinct models to incorporate Market Data (MD), Sentiment Data (SD) and Comparison Parameters Data (CPD) (see Figure 4.2).

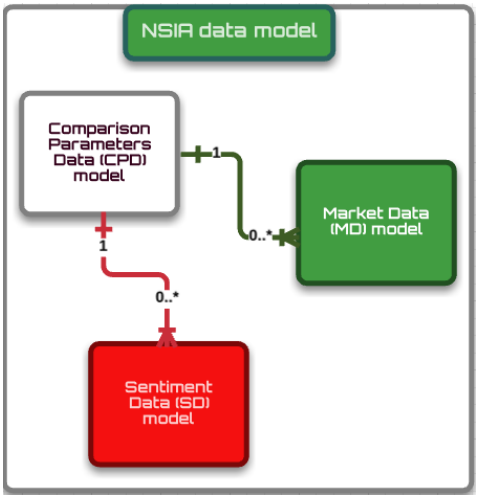


Figure 4.2 NSIA data model components

4.2.1 Market Data (MD) model

The Market Data model represents the datasets provided by financial market data providers. This model is generic and flexible enough to capture any dataset originating from providers such as Thomson Reuters (2014) and Bloomberg (2016). The conceptual data model adopted from (Rabhi et al., 2009; Milosevic et al., 2016) and presented in Figure 4.3 is composed of number of *entities*, which are described as follows:

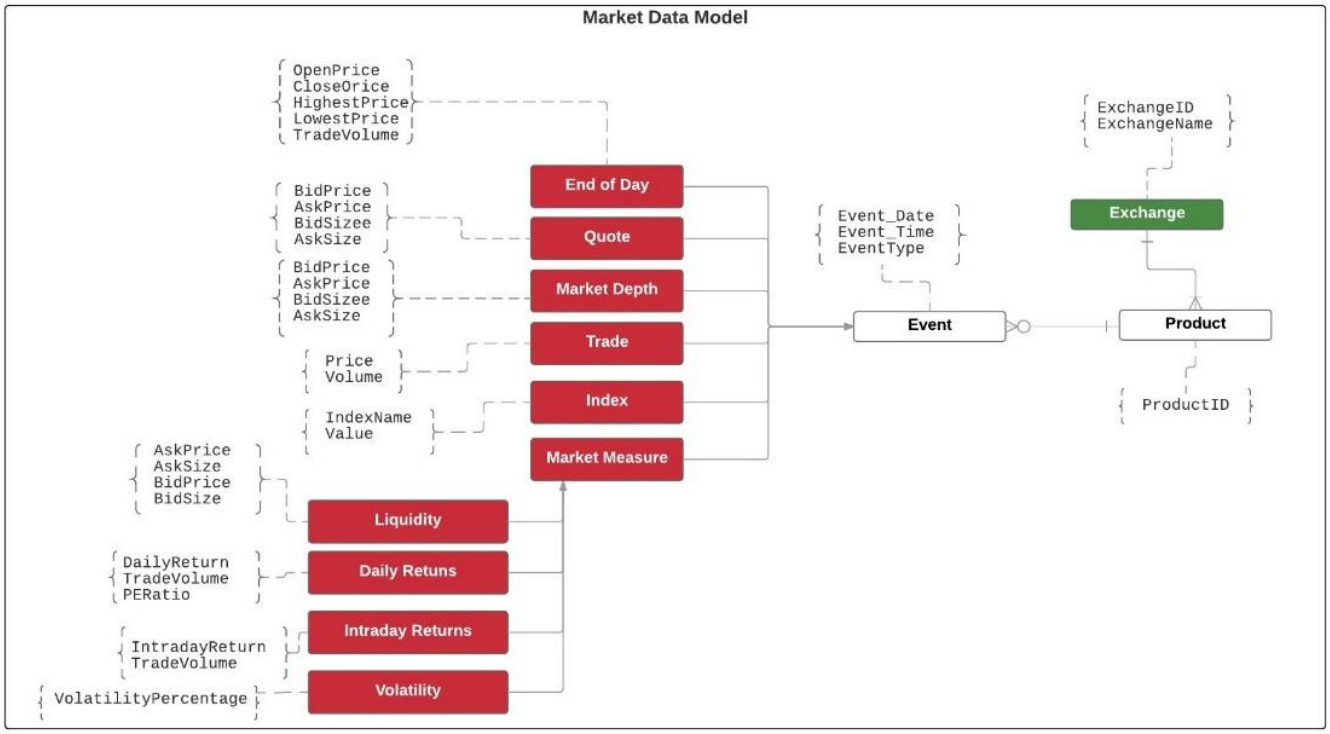


Figure 4.3 Market Data Conceptual model

- *Event*: time-stamped superclass capturing different types of events as shown in Figure 4.3. It could be extended to represent any event across different domains, for example news events.
- Product: Products are distinguished by *ProductID* key. There are two types of products: tradable and non-tradable products. Trades, quotes, end of day and market depth are tradable events, their *ProductID* would be the code of the company issuing these products. Non-tradable products include index, news and measure events. An index event's *ProductID* would be the index code.
- *Exchange*: provide platforms to trade products. Companies list and trade their products on the exchange. Exchanges maintain market datasets in either high-frequency form (Tsay, 2005), or in low-frequency form as in end of day transactions.

The market data model defines End of Day, Quote, Trade, Market Depth, Index and Market Measure events. These are described as follows:

- *End of Day*: timestamped events that represent values of trades on daily basis. Those include: Opening Price, Closing Price, Highest and Lowest Price values for a particular *ProductID.*

- *Quote*: timestamped event that lists the best bid and ask submitted to the exchange by market participants (brokers, traders) for a particular *ProductID*.
- *Trade*: timestamped events, which show the trades that took place, for a particular *ProductID*.
- *Index*: timestamped events that represent the value of a particular index.
- *Market Depth*: timestamped events showing the depth and breadth of quotes events up to a certain level (e.g. 10^{th} best bid and 10 best ask for a particular *ProductID*).
- *Market Measure*: market measure events store timestamped data related to different measures for a particular *ProductID* such as: *Liquidity, Volatility, Intraday Returns* and *Daily Returns.*

4.2.2 Sentiment Data (SD) model

The sentiment data model is designed to represent news and sentiment datasets (see subsection 2.2.4). The proposed model extends the event superclass with two events *News Items* and *News Analytics* as shown in Figure 4.4. The following is a description of these entities:

- *News Item*: timestamped event that represents a news story, which is issued on a scheduled or unscheduled basis. News items store information such as news headline, news keywords, news topics, news body and news release date and time.
- *News Analytics*: timestamped news sentiment record. This is another type of event that carries additional information about the news record. For instance, it stores sentiment related information, the news novelty and news relevance scores.

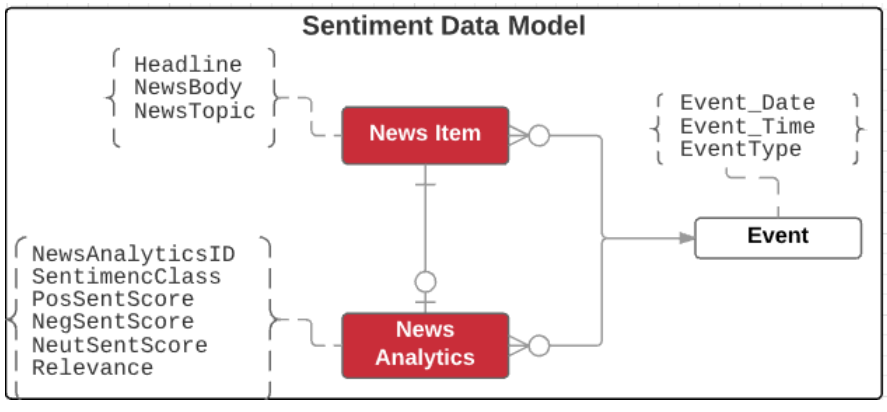


Figure 4.4 Sentiment Data conceptual model

4.2.3 Comparison Parameters Data (CPD) model

The CPD model is the key component of the NSIA data model as it represents contextual parameters associated with conducting sentiment-driven impact analysis studies.

4.2.3.1 Defining the CPD model

The Comparison Parameters Data (CPD) model divides the contextual parameters into three sets: Financial market context parameters (FC), Sentiment extraction parameters (SN) and Impact Measure parameters (IM). The financial market context parameters, shown in Table 4.1, allow the entities being impacted by the news to be defined as well as which variable is representative of the evaluation metric to be used.

Table 4.1 Financial context parameters (FC)

Parameter Name	Definition	Example
Entity E	Entity being impacted by the news	Company, an industry sector, the economy of a country as a whole.
Entity Variable Ev	Variable associated with the entity in question whose value is impacted	Closing share price, an index, GDP etc.
Benchmark B	Benchmark against which the impact will be measured	List of companies, an industry sector, the economy of a country as a whole.
Benchmark Variable Bv	Value indicative of the selected benchmark	Closing share price, an index, GDP etc.
Study Period P	The period during which the evaluation takes place	Days, Months, years…etc.

The SN parameters define sentiment extraction parameters as illustrated in Table 4.2. Besides the datasets, there are two algorithms. The first one is a filtration function, which allows a subset of the news sentiment dataset to be selected. This way, the user can decide to include, or exclude a particular category of news. For example, Thomson Reuters news include diary entries in periodic summaries, which are unlikely to have an impact so a user can decide to exclude these news from the evaluation study. The second algorithm selects the news sentiment records whose attributes values are considered to be “extreme” i.e. sentiment records whose impact will be analysed. For example, most sentiment datasets provide multiple attributes such as news relevance, news sentiment score, news sentiment class, but they can be aggregated in different ways to decide which ones to consider for impact analysis.

Table 4.2 Sentiment extraction parameters (*SN*)

Parameter Name	Definition	Example
Sentiment Dataset (M)	Name of sentiment dataset being evaluated	AlchemyAPI(2016) ,Lexalytics (2016), Quandl (2016), Thomson Reuters (2014) and RavenPack News Analytics (2016)
Filtration function (FA)	Selecting records of interest based on the attributes (fields) of a news sentiment record denoted as {a1, a2⋯an}	(Sentiment class = positive), (sentiment score > 0), (news type = Alert) ⋯etc.
Extreme Sentiment Extraction (ESE) Algorithms	Ranking algorithms, that define the basis for selecting "extreme" news sentiments	For example, extract top 5% negative news records

Finally, the Impact Measure (IM) parameters are used to measure the impact of the news sentiment scores for a given set of Financial Context (FC) parameters as shown in Table 4.3. Impact Measure (IM) parameter can be represented as one of the different financial impact measures to be discussed in subsection 4.2.3.4. Depending on the IM parameters selected, some impact measures require the user to set the Estimation Window (EW) parameter. This parameter enables the user to set the period of estimation, to calculate the expected (predicted value) of a particular measure, which enables the CPD model to compare between the predicted and the actual impact figures and the user to understand the impact magnitude of the "extreme" news sentiment records identified by the SN parameters.

Table 4.3 Impact measure parameters (*IM*)

Parameter Name	Definition	Example
Impact Measure Parameter IM	Specifies how to measure the impact of news sentiment on the entity (E), relative to the benchmark (B)	Daily Mean Cumulative Average Abnormal Returns, Intraday Mean Cumulative Average Abnormal Returns …etc.
Estimation Window EW	The estimation period used to measure impact	Hours, Days, Months …etc.

A high-level overview of the CPD model is provided in Figure 4.5. These sets of parameters are mapped to entities and attributes of the *Market Data* model and the *Sentiment Data* model. Each impact study performed by the user is distinguished from other studies using a unique key *StudyNo* attribute. An abstract instance of CPD model would be composed of attributes defined as: {Study No, Financial Context (FC) parameters, Sentiment Extraction (SN) parameters, Impact Measure (IM) parameters}. We now describe the entities comprising each set of parameters in more detail.

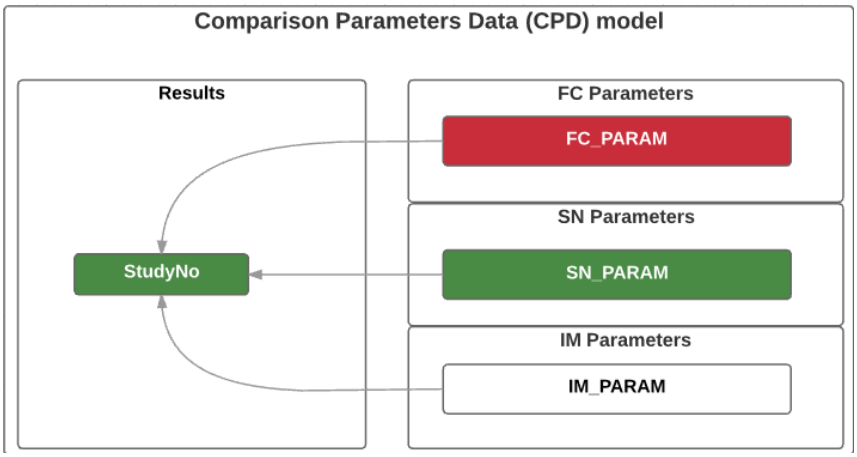


Figure 4.5 High level view of CPD model

4.2.3.2 Financial Context (FC) parameters

The *FC_PARAM* entity represents a given financial context and is linked to entities in the market data model through the following relationships (see Figure 4.6):

- *CtxEntity*: This relationship links financial entities (such as company ids) that are part of the impact study to the corresponding *Product* entity as shown in Figure 4.6.

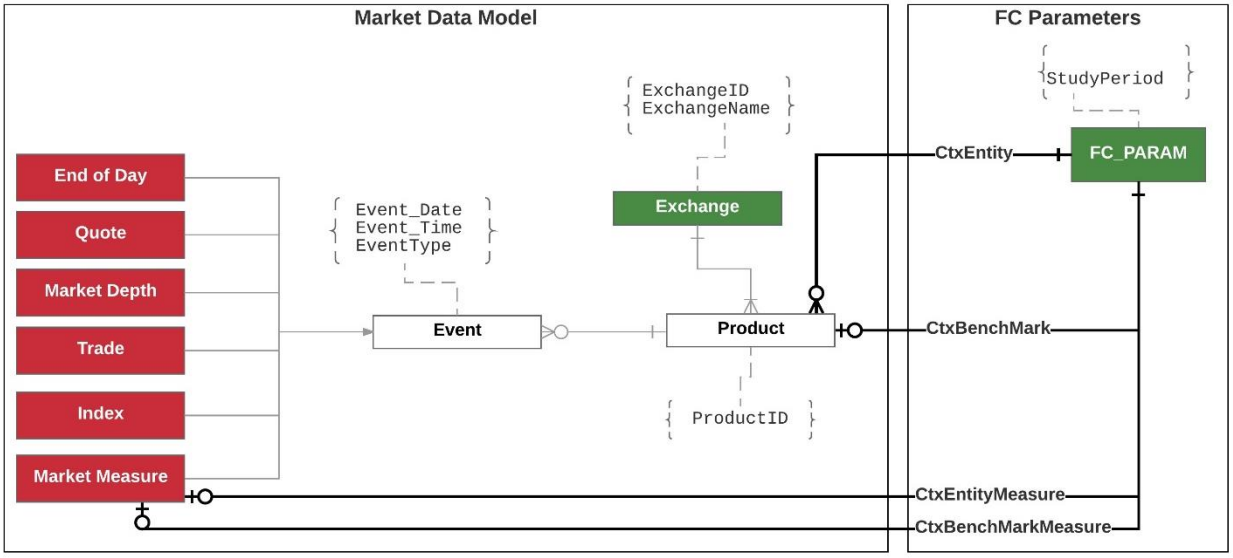


Figure 4.6 Defining Financial Context model

- *CtxEntityMeasure*: This relationship enables associating the financial context with a *MarketMeasure* event. A measurable event could be intraday return, daily returns …etc.

- *CtxBenchmark*: This relationship defines a benchmark that is offered as a product in the *Product* entity, usually an *Index* event. Defining benchmark entities is a method used in many event data evaluation studies (Bohn et al., 2012).
- *CtxBenchmarkMeasure*: This relationship associates a benchmark with a market measure event in the *MarketMeasure* entity. For instance, it could represent an aggregate value of an event, say for example a business sector value, an interest rate value, or an index value.

The *FC_PARAM* entity has the following attributes:

- *StudyPeriod*: This attribute represents the date or time ranges of events. It is mapped to *EventDate* and *EventTime* attribute in *Event* entity.

4.2.3.3 Sentiment Extraction parameters (SN)

The SN parameters in the CPD model define five entities as shown in Figure 4.7, to filter news and identify sentiment datasets relevant for the study (see Table 4.2). These are explained as follows:

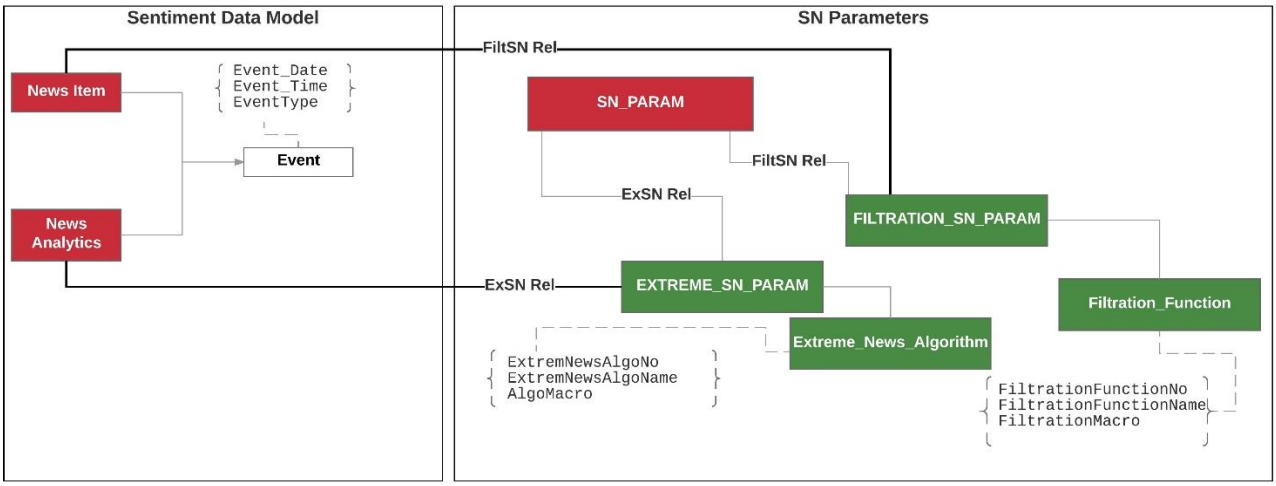


Figure 4.7 Defining SN parameter entities

- *SN_PARAM*: the main entity which acts as a root for the other entities in the SN parameters model.
- *FILTRATION_SN_PARAM* and *Filtration_Function*: the first entity links to the *News Item* instances via the relationship *FiltSN Rel*. These instances are produced by the function defined in the *Filtration_Function* entity.

- *EXTREME_SN_PARAM* and *Extreme_News_Algorithm*: the first entity links to the *News Analytics* instances via the relationship *ExtSN Rel.* These instances are produced by the algorithm defined in the *Extreme_News_Algorithm* entity.

4.2.3.4 Impact Measure parameters (*IM*)

The CPD model defines *IM_PARAM* and *IMPACT_MEASURE* entities, which enable the CPD model to apply different impact models (see Figure 4.8).

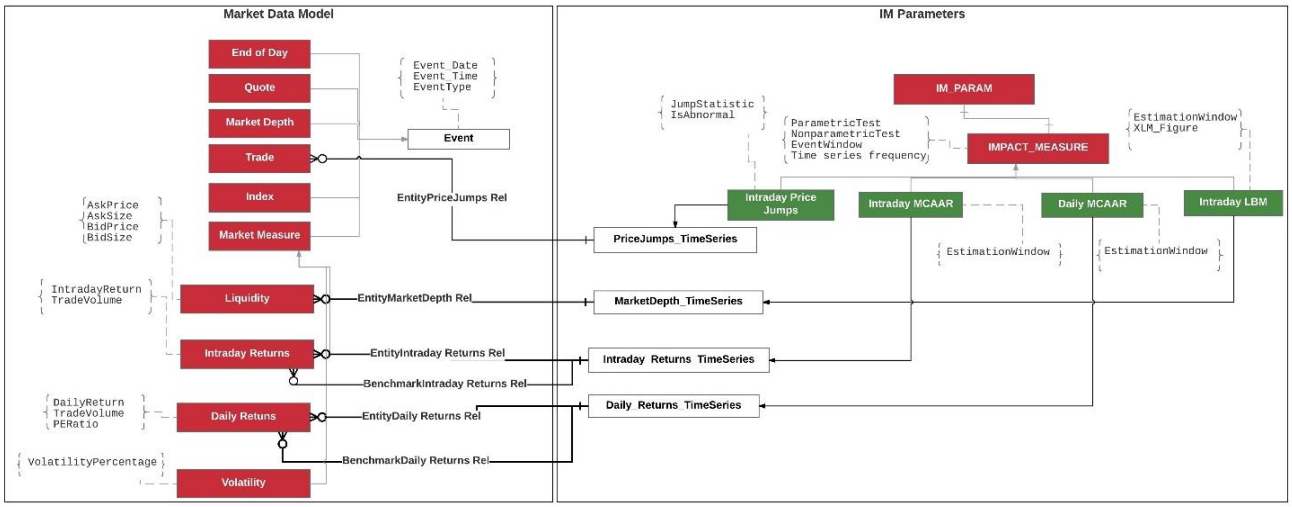


Figure 4.8 Defining Impact Measures parameters

These entities are defined as follows:

- *IM_PARAM*: the root entity which connects the *StudyNo* with the impact measure used in that study.
- *IMPACT_MEASURE*: this is a superclass which facilitates the implementation of various impact models, as discussed in subsection 4.2.3.1.

In this thesis, we choose to provide four impact measures that illustrate the variety of impact models that could be represented in the CPD model. These four impacts measures are:

- Daily Mean Cumulative Average Abnormal Returns (Daily MCAAR): this impact measure uses an event study methodology explained in (Agrawal et al., 2006), to calculate the daily mean cumulative average abnormal returns.

- Intraday Mean Cumulative Average Abnormal Returns (Intraday MCAAR): this measure is applied to intraday data. It uses high frequency returns time series to calculate the intraday (e.g. 5-minute, 10-minute intervals) mean cumulative average abnormal returns methodology as demonstrated in (Siering, 2012a).
- Intraday Price Jumps: measures the volatilities in stock price timeseries data. The measure used based on the method proposed in (Lee & Mykland, 2007) and used in (Bohn et al., 2012), which is capable of capturing the timing and size of price jumps. The measure applies a threshold over the stock prices observations and a price jump is recorded if an observation breaks through the threshold.
- Intraday Liquidity Based Model (Intraday LBM): this measure uses the EXchange Liquidity Measure (XLM) method, which calculates the trading costs of a roundtrip trade of a given size as explained in (Gomber et al., 2015).

Accordingly, a number of time series entities which preserve timeseries data calculated using these impact models, are defined as subclasses of the *IMPACT_MEASURE* entity. These entities are:

- *Daily_Returns_TimeSeries*: This entity preserves instances of Daily Returns dataset via EntityDaily Returns Rel and BenchmarkDaily Returns Rel relationships.
- *Intraday_Returns_TimeSeries*: This entity preserves instances of Intraday Returns dataset via EntityIntraday Returns Rel and BenchmarkIntraday Returns Rel relationships.
- *PriceJumps_TimeSeries*: This entity preserves instances of *Trade* dataset via *EntityPriceJumps Rel* relationship.
- *MarketDepth_TimeSeries*: This entity preserves instances of *Liquidity* dataset via *EntityMarketDepth Rel* relationship.

4.2.3.5 CPD model results

The CPD model logs the set of parameters selected for each impact study using the *CPD_PARAM_LOG* entity, distinguishing studies by *StudyNo* attribute.

4.3 NSIA architecture

4.3.1 Overview of NSIA architecture

The NSIA architecture (shown in Figure 4.9) is designed to support the proposed data model and the impact analysis use cases. The NSIA architecture follows the ADAGE Framework guidelines (Rabhi et al., 2012) and is designed using a combination of both component based and service oriented design principles. The architecture encompasses three layers: User layer, Business layer and Data layer. These layers are summarized as follows:

- GUI layer: mediates the interactions between users and the Business layer, based on user selections. The user interfaces provided in the NSIA architecture enable actionizing the data model defined in section 4.2. Users through the GUI layer invoke use cases defined in section 4.4.
- Business layer consolidates a number of components which encapsulate the majority of the framework's business logic.
- Data layer: a number of data repositories are used to cater for the complete cycle of conducting sentiment-driven impact analysis studies.

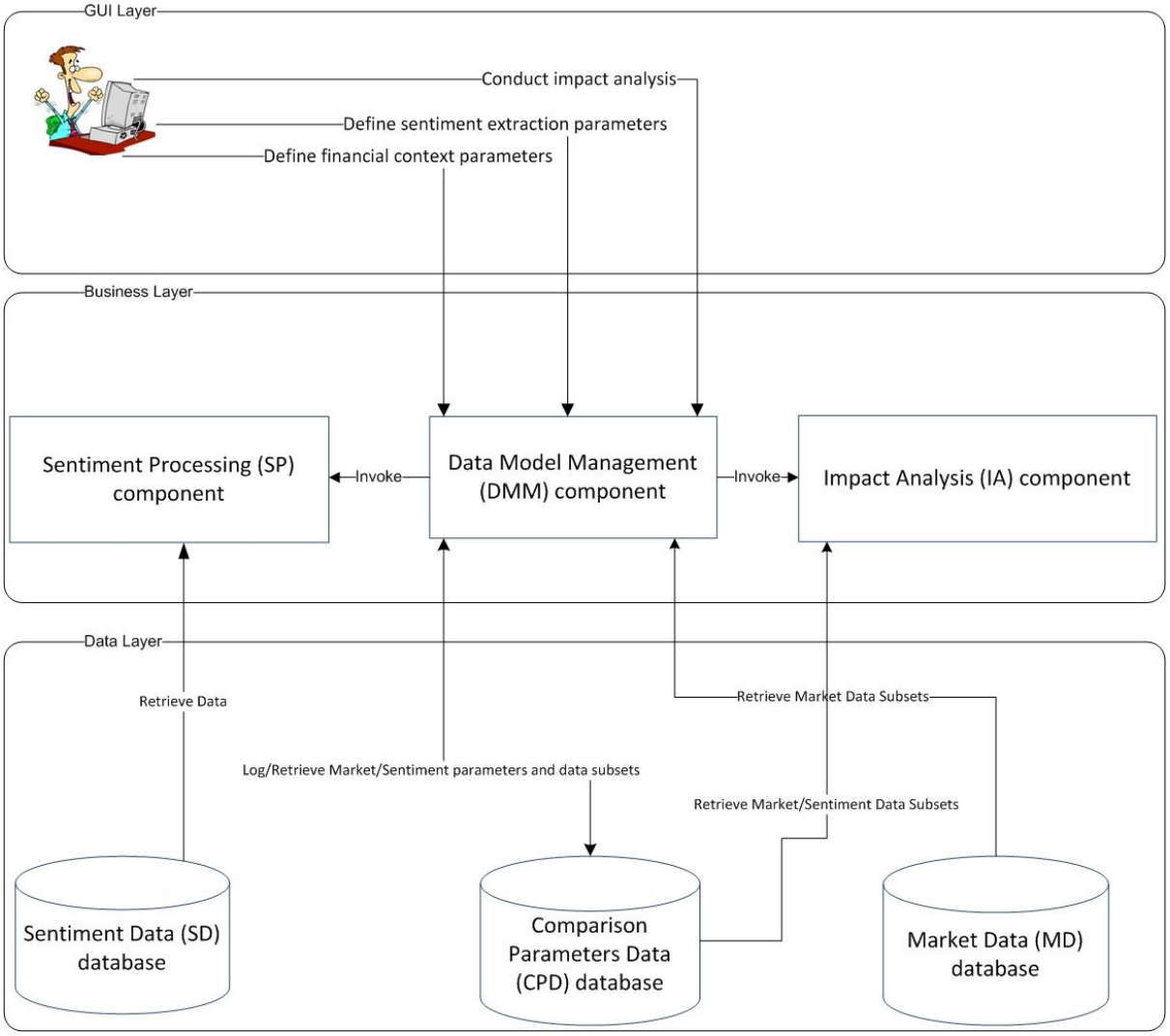


Figure 4.9 NSIA architecture

4.3.2 Business layer

The NSIA Business layer consists of number of components, these are:

- Data Model Management (DMM) component: this component is in charge of creating, accessing, and updating all the entities in the data model described in section 4.2.
- Sentiment Processing (SP) component: this component is utilized to extract data according to the SN parameters (see subsection 4.2.3.3).
- Impact Analysis (IA) component: this component enables the user to undertake different impact studies, based on the impact models defined in subsection 4.2.3.4.

The role of these components in implementing the use cases will be described in the next section.

4.3.3 Data layer

This layer preserves the conceptual data models presented in section 4.2 into a data preserving mechanism e.g. databases. There are three data repositories created to cater for this purpose. All parameter sets, and data needed to conduct an impact analysis study using the NSIA framework is physically stored in this layer, these repositories are:

- Market Data (MD) database: data repository manages and preserve the market data model entities defined in subsection 4.2.1.
- Sentiment Data (SD) database: data repository manages and preserve the sentiment data model entities defined in subsection 4.2.2.
- Comparison Parameters Data (CPD) database: This database responsible of managing all the entities related to the Comparison Parameters Data (CPD) model, depicted in subsection 4.2.3.

4.4 NSIA use cases

The third design artefact introduces a number of *use cases* to guide the user in conducting sentiment-driven impact analysis studies using the NSIA data model and architecture.

4.4.1 Overview of use cases

Figure 4.10 shows the three use cases, which are:

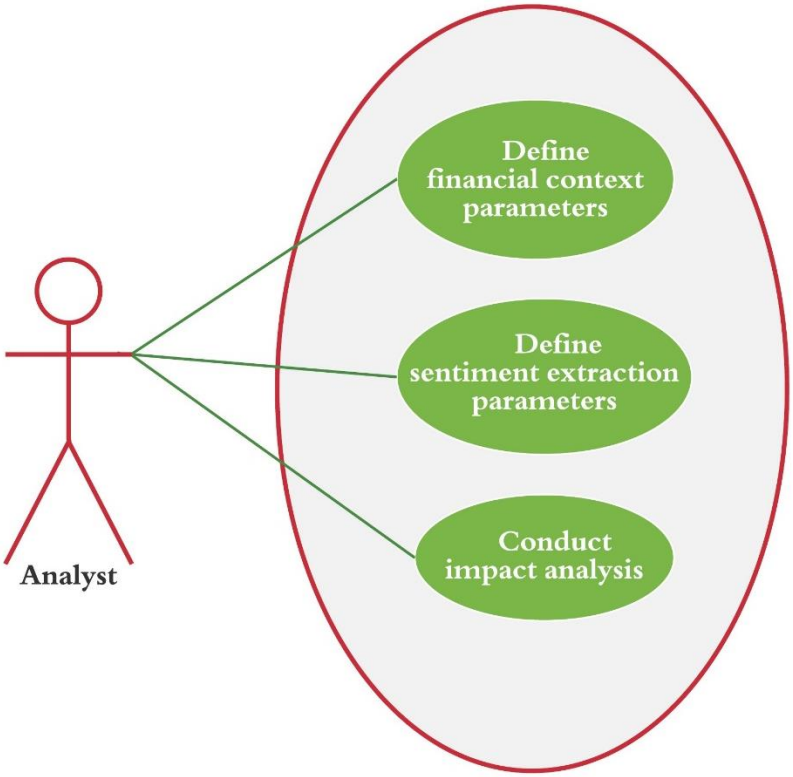


Figure 4.10 Use case to define CPD parameters and execute impact analysis studies

- Define financial context parameters: this use case guides the analyst/user in defining the financial context parameters (see subsection 4.2.3.2).
- Define Sentiment extraction parameters: this use case assists the analyst/user in defining the sentiment extraction parameters (see subsection 4.2.3.3).
- Conduct impact analysis: this use case assists the user in defining the Impact Measure parameters (see subsection 4.2.3.4) and conducting impact analysis studies.

The sequence diagrams that correspond to these three use cases are now described in more detail in the rest of the section.

4.4.2 Define financial context parameters

The use case sequence diagram is shown in Figure 4.11. Users can set the FC parameters via a GUI and load the market data needed for the evaluation. The Data Model Management (DMM) component is responsible for logging the FC parameters in the CPD database, generating a study number for the

new impact study and for creating market data subsets relevant to the FC parameters, then log them to the CPD database.

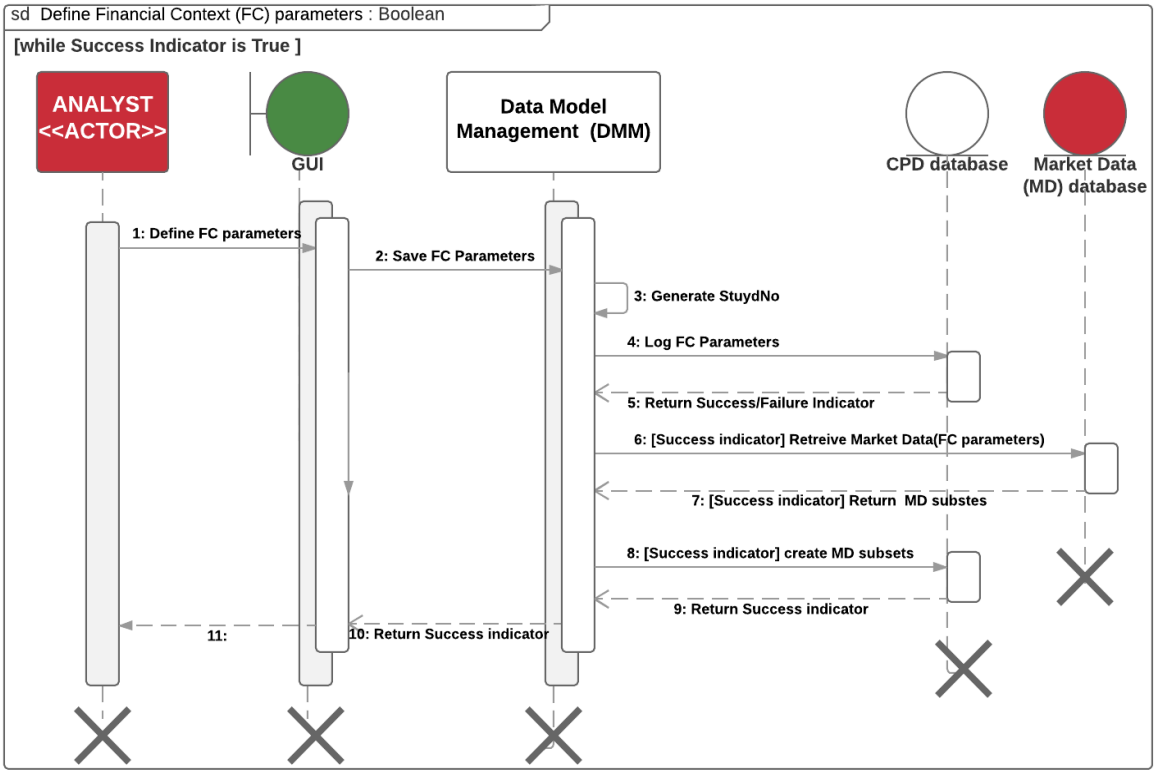


Figure 4.11 Define financial context parameters sequence diagram

4.4.3 Define sentiment extraction parameters

The use case sequence diagram is shown in Figure 4.12. Users can set the SN parameters via a GUI and load the sentiment data needed for the evaluation. The Data Model Management (DMM) component is responsible for logging the SN parameters in the CPD database (see subsection 4.2.3.3) and invoking the Sentiment Processing (SP) component, which filters news sentiment (according to Filtration function FA) and creates subsets of extreme news sentiment datasets (according to the Extreme Sentiment Extraction (ESE) algorithm) in the CPD database.

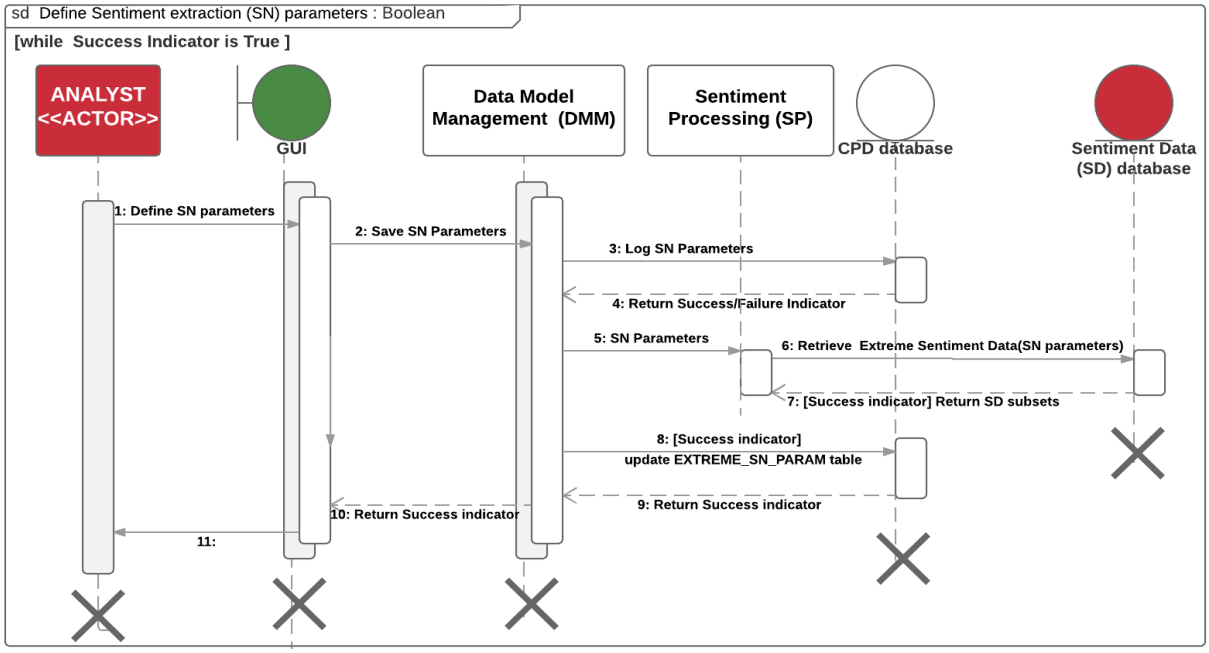


Figure 4.12 Define Sentiment extraction (SN) parameters sequence diagram

4.4.4 Conduct impact analysis

The impact analysis use case assumes that FC and SN parameters have been defined and that market and sentiment datasets, which are the subject of the evaluation have been identified. The use case sequence diagram is shown in Figure 4.13. It starts when the user defines the appropriate Impact Measure (IM) parameters, which are saved in the CPD database by the DMM component. The user then invokes Impact Analysis (IA) component via a GUI, which retrieves the relevant market and sentiment subsets as per step 6 in Figure 4.13. The IA component applies statistical significance tests to compute the impact results as per step 8. The IA component then logs the impact results to the CPD database. Some examples will be described in chapter 5.

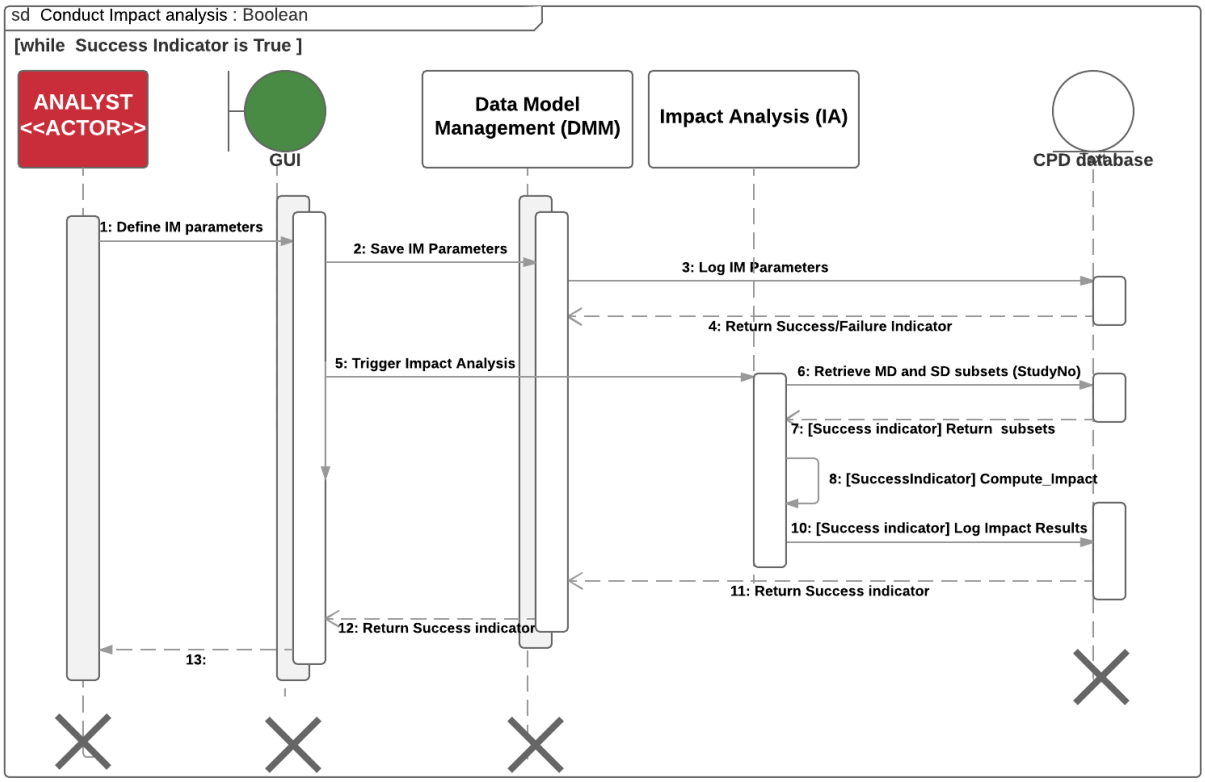


Figure 4.13 Conduct Impact analysis sequence diagram

4.5 Conclusion

In this chapter, we introduced the NSIA framework's three essential components. The first one is a Data model, which provides the flexibility to integrate a variety of sentiment and market datasets. It also enables the user to define through the Comparison Parameters Data(CPD) model three sets of parameters, which are relevant to the impact analysis evaluation. The second one is a number of architectural components that provide various functions for defining parameters, creating subsets for evaluation and carrying out impact analysis. The third one are use cases to guide the user in conducting sentiment-driven impact analysis studies.

The design of the framework constitutes the second stage in the research process, which is the design stage (see subsection 3.4). The next chapter, is dedicated to the implementation stage and will describe a prototype of the NSIA framework that has been developed to provide a basis for the evaluation stage.

5 PROTOTYPE IMPLEMENTATION

This chapter describes a prototype implementation of the NSIA framework proposed in chapter 4. The goal behind the implementation is to validate the functionality and feasibility of the proposed framework. The NSIA data model is implemented in section 5.1, the software architecture implemented in section 5.2 and the use cases are described in section 5.3. The chapter describes some of the limitations of the porotype in section 5.4. Finally, the chapter concludes in section 5.5.

5.1 Implementing the Data layer

The NSIA data model presented in section 4.2, proposes a number of *entities* that capture concepts related to conducting sentiment-driven impact analysis studies. In this prototype implementation, we first describe the datasets used, explain the implementation choices that have been made, then describe how the three data models (Sentiment Data model, Market Data model and Comparison Parameters Data model) have been implemented.

5.1.1 Datasets used

The datasets used in this prototype have been acquired from Thomson Reuters (2014) through Sirca (https://www.sirca.org.au/) which provides subscribing universities in Australia access to financial data repositories. In particular, the prototype used the following two datasets:

- Thomson Reuters Tick History (TRTH): consists of high frequency data such as Trades, Quotes, Market Depth occurrences. Also, it provides market data with lower frequency for End of Day prices. Datasets were acquired for selected companies in the period 2003 to 2011.
- Thomson Reuters News Analytics (TRNA): contains over 7 million news analytics records related to over 20000 companies trading on exchanges from all around the world. Most news are of finance nature covering stories from 2003 to 2011.

Both these datasets organize data in the form of textual files included in csv format, to make it easier for finance researchers to perform further analysis using statistical packages. To populate the *Trade, Quote*, *End of Day* and *Market Depth* entities, this implementation utilized a web portal provided directly through Sirca (2017) as shown in Figure 5.1 and Figure 5.2. This web portal expects two parameters a list of Reuters Instrument Codes (RIC) and a date range, the data delivered by Thomson Reuters as csv files.

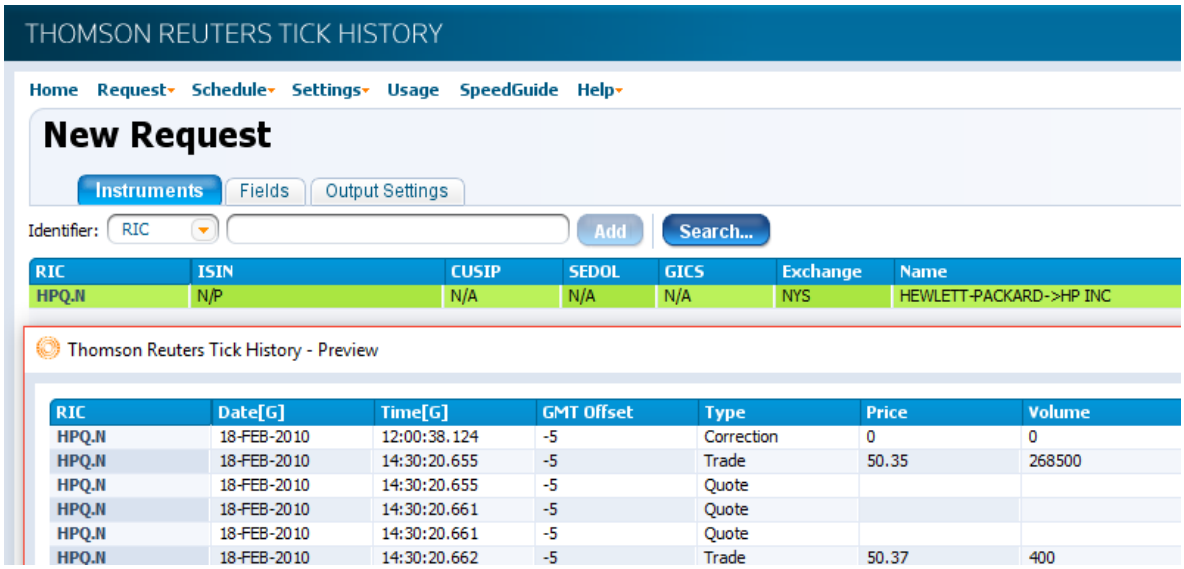


Figure 5.1 Thomson Reuters Tick History web portal showing trades and quotes

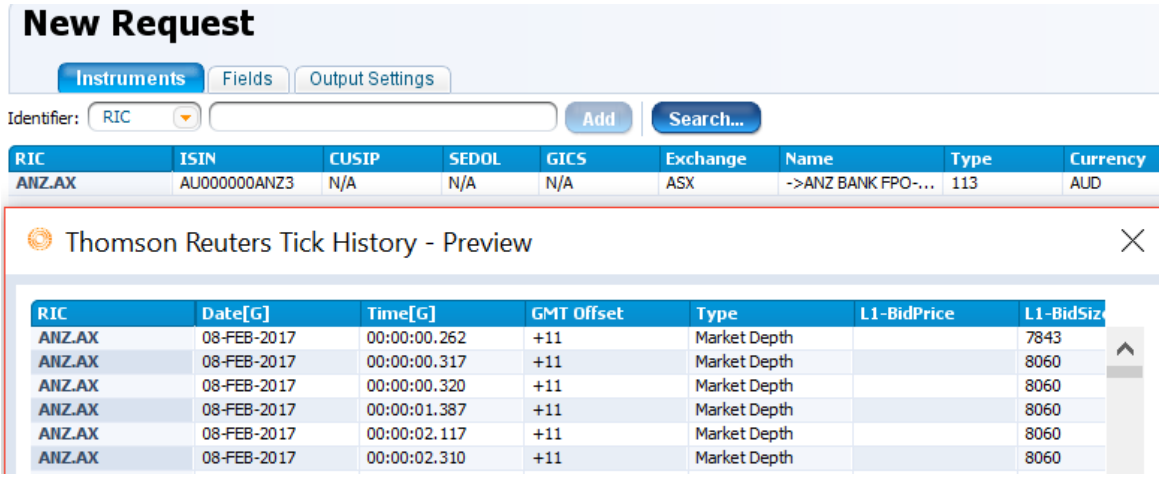


Figure 5.2 Thomson Reuters Tick History web portal showing Market Depth data

5.1.2 Implementation Choices

The Data layer in the prototype has been implemented mostly using an Oracle database 11g. The motivation behind choosing a relational database to store some of the important entities is due to the following reasons (Hesham, 2017):

- Fast response to information requests: querying database is much faster than looking up data in a sheet, especially if there are multiple sheets involved.

- Flexibility: queries submitted to database could be very sophisticated, and efficient. The same functionality is hard to implement with traditional data files i.e. excel sheets.
- Less storage: databases require less storage, as the data is stored once, while with file systems data could be redundant, and end up taking much more space on the desk.

However, due to the size of Thomson Reuters datasets used, we made the following implementation choices:

- Raw data were kept in csv files.
- Timeseries data are stored in database objects.

To populate the database objects *Oracle SQL developer* utility was used (see Figure 5.3).

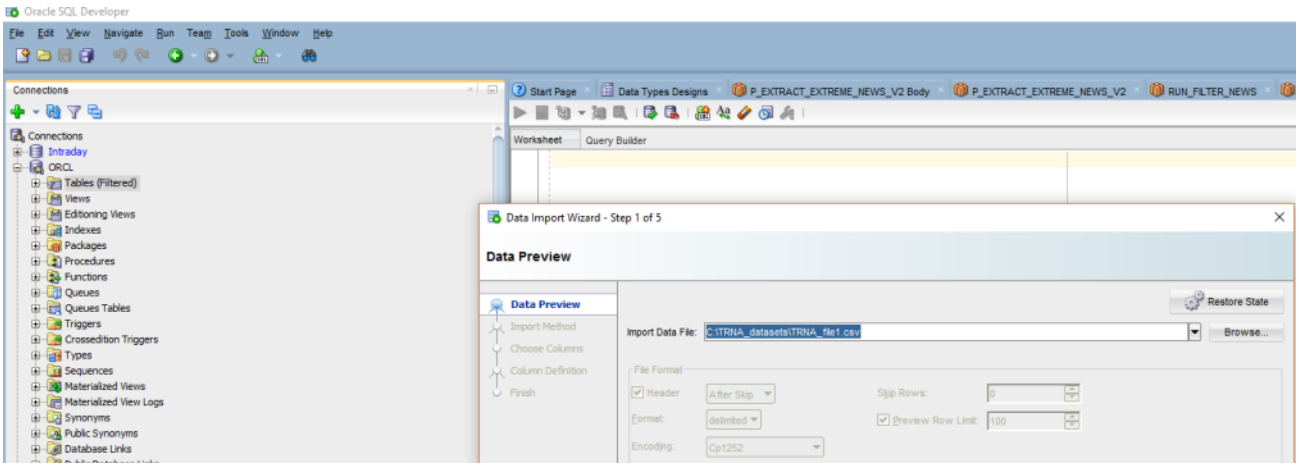


Figure 5.3 Oracle SQL developer import data utility (Oracle Corporation, 2017a)

5.1.3 Implementing the Market Data (MD) model

The entities in the Market Data model described in subsection 4.2.1 have been implemented as follows:

- The primary key for *Exchange* entity (*Exchange ID*) follows Thomson Reuters naming convention. For example, ASX will be represented by the code AX and N= New York Stock Exchange.
- The primary key for *Product* entity (*Product ID*) is represented by a Thomson Reuters Instrument Code (RIC). RICs are structured codes used to uniquely identify any financial instrument such as stocks and indices in all its data products. For example, the Reuters Code for Hewlett Packard

listed on New York Stock Exchange is HPQ.N. Market indices are prefixed with a dot in front of their RIC symbol, SPX for S&P 500 or. DJI for Dow Jones Industrial Average.

Therefore, instances of the *Event* entity will be stored in two different ways depending on the subclass concerned. As Table 5.1. shows, the Market Data model entities: *Trade*, *Quote*, *Market Depth* and *End of Day* have been stored as CSV files. Other Market Data model entities such as *Index* and *Market Measure* entities (*Liquidity*, *Daily Returns* and *Intraday Returns*) are represented using the Oracle database objects shown in Table 5.1.

Table 5.1 Mapping the Market Data model entities to Database objects

Market Data model entity	Mapped prototype object	Storage Type	Data type
Exchange	GLOB_ENTITY	Database object	Timeseries data
Product	GLOB_ENTITY	Database object	Timeseries data
Trade	TRTH format	CSV files	Raw data
Quote	TRTH format	CSV files	Raw data
Market Depth	Market Depth TRTH format	CSV files	Raw data
End of Day	TRTH format	CSV files	Raw data
Index	INTRADAY_HOMO_TS_INDEX	Database object	Timeseries data
Liquidity	INTRADAY_MARKETDEPTH_XLM 50000	Database object	Timeseries data
Daily Returns	DAILY_RETURNS	Database object	Timeseries data
Intraday Returns	INTRADAY_HOMO_TS	Database object	Timeseries data

In the prototype, all database table structures are created using Oracle scripts, an example of which for the database object INTRADAY_HOMO_TS is shown in Figure 5.4.

```
CREATE TABLE "INTRADAY_HOMO_TS"
   (   "SECURITY" VARCHAR2(260 BYTE),
       "DATEOFTRADE" DATE,
       "STARTPERIOD" VARCHAR2(26 BYTE),
       "ENDPERIOD" VARCHAR2(26 BYTE),
       "LSTPRICE" NUMBER (25,15),
       "LSTRETURN" NUMBER (38,30),
       "TOTVOL" NUMBER (10,0));
```

Figure 5.4 Creates INTRADAY_HOMO_TS table structure

To create the *Index*, *Liquidity*, *Intraday Returns* and *Daily Returns* object instances (see Table 5.1), the Haskell Generic Aggregator(HGA) tool was used (Rabhi et al., 2012; Yao and Rabhi, 2015). This tool creates and merges time series data which requires some parameters to be defined. It produces timeseries files as shown in Table 5.2 in CSV file format. A combination of Unix shell scripts (Stonebank, 2000) and Oracle SQL Loader utility (Oracle Corporation, 2017c) was used, to load the csv files into the prototype database objects. The Unix shell script is shown in Appendix A (Figure A. 3).

Table 5.2 HGA role in producing timeseries csv files

Input Market Data entity	**Input type**	**Output type**	**Input/ Output File type**	**Prototype object**
End of Day	TRTH format	Daily timeseries	CSV file	DAILY_RETURNS
Trade and Quote	TRTH format	Intraday homogenous timeseries	CSV file	INTRADAY_HOMO_TS
Index	TRTH format	Intraday index homogenous timeseries	CSV file	INTRADAY_HOMO_TS_INDEX
Market Depth	Market Depth TRTH format	Intraday Market Depth homogenous timeseries	CSV file	INTRADAY_MARKETDEPTH_XLM50000

In Figure 5.5 an example of CSV output produced by HGA tool is shown of a sample intraday timeseries returns occurrences for Siemens Corporation.

SECURITY	DATEOFTRADE	STARTPERIOD	ENDPERIOD	LSTPRICE	LSTRETURN	TOTVOL
SIEGn.DE	08/MAR/10	09:35:00	09:40:00	67.42	-0.001037729	17135
SIEGn.DE	08/MAR/10	09:40:00	09:45:00	67.35	-0.001038807	19672
SIEGn.DE	08/MAR/10	09:45:00	09:50:00	67.37	0.000296912	3462
SIEGn.DE	08/MAR/10	09:50:00	09:55:00	67.3	-0.001039578	13468
SIEGn.DE	08/MAR/10	09:55:00	10:00:00	67.29	-0.000148599	19021

Figure 5.5 Sample Intraday returns for German company Siemens

5.1.4 Implementing the Sentiment Data (SD) model

The Sentiment Data (SD) model entities *News Analytics* and *News Item* (described in subsection 4.2.2) have been implemented using a single *database object* called SENT_RAW_DATA. This database object consists of 87 attributes, capturing information related to the news meta data, such as news type, relevance score to *Product* entity, novelty attributes, and sentiment scores attributes, as well as dates and times of news release. Figure 5.6 shows the mapping between the proposed entities in the sentiment data model and the attributes in SENT_RAW_DATA database object.

News Analytics & News Company Link	SENT_RAW_DATA	Example
NewsID	PNAC	nSYU009933
NewsAnalyticsID	ITEM_ID	rna.service1.2010-05-17_00.27.36.nSYU009933.A1.0faedebe
ReleaseDate	STORY_DATE	17 MAY 2010
ReleaseTime	STORY_TIME	00:27:36
NewsType	ITEM_TYPE	ALERT
SentimentClass	SENTIMENT	-1
SentimentScore	SENT_SCORE	0.81913
Novelty	LNKD_CNT5	17
NewsHeadline	BCAST_TEXT	BHP BILLITON <BHP.AX> WARNS PROPOSED NEW TAX WILL ALSO HI
NumberWords	TOT_WORDS	13
NewsTopic	TOPIC_CODE	AU ASIA BMAT MEMI MINE MTAL MIS MET ZA GB CN NEWS US CA IRN
CompanyID	STOCK_RIC	BHP.AX

Figure 5.6 Mapping SD entities to SENT_RAW_DATA database object

5.1.5 Implementing the Comparison Parameters Data (CPD) model

The entities described in subsection 4.2.3 related to the Comparison Parameters Data (CPD) model have been implemented using database objects, as defined in Table 5.3.

Table 5.3 Mapping CPD objects and the Implementation objects

CPD Model	CPD model entity	Mapped prototype object
Financial Context (FC) parameters	*FC_PARAM*	IMPACT_STUDIES_LOG
Sentiment Extraction parameters (SN)	*SN_PARAM*	IMPACT_STUDIES_LOG
	FILTRATION_SN_PARAM	TRNA_SCORES
	EXTREME_SN_PARAM	TRNA_SCORES_APPLIED
	Extreme_News_Algorithm	TRNA_EXTREME_ALGO
	Filtration_Function	TRNA_FILTRAION_FUNC
Impact Measure (IM) parameters	*IM_PARAM*	IMPACT_STUDIES_LOG
	Intraday_Returns_Timeseries	MERGE_TSLOT_AR
	Daily_Returns_Timeseries	DAILY_MCAR_DATA
	MarketDepth_Timeseries	MERGE_MARKETDEPTH_TIME SLOTS
	PriceJumps_Timeseries	*MERGE_PRICEJUMPS_TIMESLO TS*
	Intraday Price Jumps	*IMPACT_RESULTS*
	Intraday MCAAR	*IMPACT_RESULTS*
	Intraday LBM	*IMPACT_RESULTS*
	Daily MCAAR	*IMPACT_RESULTS*

Appendix A (Figure A. 1) portrays a sample of IMPACT_STUDIES_LOG database object that defines CPD parameters for some of the studies conducted in chapter 6. As another example, Appendix A (Figure A. 2) defines the physical structure of INTRADAY_PRICE_JUMPS database object. Attributes *JUMPSTATISTIC*, *ISABNORMAL* and *NEWS_ABNORMAL* store the price jump statistic relevant to the intraday time series record.

The timeseries objects created in the CPD model (see subsection 4.2.3.4), are implemented using the following objects:

- MERGE_TSLOT_AR database object: SQL object that implements intraday impact logic to calculate intraday abnormal returns data.
- DAILY_MCAR_DATA database object: used as storage of applying daily abnormal returns.
- MERGE_MARKETDEPTH_TIMESLOTS: this database object applies aggregates market depth time series data on intervals less than a day, for example 5-minute slots.
- MERGE_PRICEJUMPS_TIMESLOTS*:* this database object applies aggregates price jumps time series data on intervals less than a day, for example 5-minute slots.

5.2 Implementing the Business layer

5.2.1 Implementation choices

The Business layer consists of a number of components: Data Model Management (DMM) component, Sentiment Processing (SP) component, and Impact Analysis (IA) component (see subsection 4.3.2). These components were implemented using a number of technologies (see Table 5.4) including Eventus (Cowan Research,2016), R software (2017), and PL/SQL database stored procedures (Oracle Corporation, 2017b). Eventus is a software tool that is based on SAS statistical analysis programming language used to compute abnormal returns of an event. R software is a free software which consolidates an extensive number of data visualization and analysis packages.PL/SQL stored procedures are powerful, efficient, and robust scripting language that enables software developers to store sophisticated data processing scripts in the database. These implementation choices are applicable to certain impact models (see subsection 4.2.3.4) as shown in Table 5.4.

Table 5.4 Mapping technologies to their corresponding implemented impact models

Software tool	**Impact Analysis (IA) component**	**Data Model Management (DMM) component**	**Sentiment Processing (SP) component**
Intraday MCAAR	R software	PL/SQL Stored procedures	PL/SQL Stored procedures
Intraday Price Jumps	R software	PL/SQL Stored procedures	PL/SQL Stored procedures

Intraday LBM	R software	PL/SQL Stored procedures	PL/SQL Stored procedures
Daily MCAAR	Eventus	PL/SQL Stored procedures	PL/SQL Stored procedures

Further detailed description of how the Business layer has been implemented is given in the rest of this section.

5.2.2 Implementing the Sentiment Processing (SP) component

The CPD model entities *Filtration_Function* and *Extreme_News_Algorithm* (see subsection 4.2.3.3) have been implemented using PL/SQL methods. These are explained as follows:

5.2.2.1 Filtration_Function method

The method defined in Figure 5.7 is a PL/SQL script implementing the logic of *FILTRATION_FUNCTION* entity, which has been described in subsection 4.2.3.3. It encapsulates a number of filtration functions, each of which has its own number (*p_Filtration_Function_No*) and filtration attributes (*p_Filtration_Parameters*). Upon invoking this method, a subset of news sentiment records is created and saved to TRNA_SCORES database object.

```
Procedure Filtration_Function    ( p_Filtration_Function_No  NUMBER,
                                   p_Filtration_Parameters   TRNA_SCORES%ROWTYPE
                                  )
```

Figure 5.7 Filtration_Function method definition

The filtration functions implemented in *Filtration_Function* method, utilizes part or all of the following attributes (found in TRNA dataset) defined in Table 5.5.

Table 5.5 Filtration attributes in TRNA dataset

Attribute Name	Attribute Code	Description
Reuters Instrument Code	*RIC*	The entity related to the study
News Relevance	*R*	An attribute which defines how relevant the news item is to the entity. Values range from (0 non-relevant) to (1 highly relevant).

News Topic	*NT*	Each news item is related to a certain topic/s, denoted as NT. Examples: AIR for air transport news, BKRT for bankruptcy related news, 'O' for oil news, 'JOB' for job strikes related news.
News Item Story Release Date	*SRD*	News story release date and time, which would become the event date if the relevant news record has been identified as an extreme news negative news
Sentiment Class	*SC*	News item sentiment orientation. TRNA classifies news into three classes (Positive with score of 1, Negative with score -1 and Neutral with score 0).
Sentiment Score	*SS*	Each news item has sentiment score ranging from 0 neutral/no sentiment to 1 extreme sentiment score.
Number of Companies	*NO_COMP (NC)*	Number of companies mentioned in the news item. Values can be ≥ 1. When value set to 1, it means the news item addresses one single company
Novelty	*LNKD_CN T5 (NV)*	Shows if the news item is novel or not. Values can be ≥ 0. Value 0 means the news item is the first release of the news story.

Any number of filtration functions can be defined. Table 5.6 shows examples of *Filtration_Function* that will be used in the impact studies conducted in chapter 6. The *RIC*, *SRD* attributes derive their values from the financial context parameters (Entity (E), and Study Period (P)).

Table 5.6 Filtration Functions (FA) with filtration attributes

Filtration Function (FA)	Filtration Attributes
Filtration Function 1	RIC = Entity (E) parameter, R=1, SRD = Study Period (P) parameter
Filtration Function 2	RIC= Entity (E) parameter, R=1, SRD = Study Period (P) parameter, NT = 'O'
Filtration Function 3	RIC= Entity (E) parameter, R=1, SRD = Study Period (P) parameter, NT = 'JOB'
Filtration Function 4	RIC= Entity (E) parameter, R=1, NC=1, NV=0, SRD= Study Period (P) parameter

5.2.2.2 Extreme_News_Algorithm method

The method implemented in an Oracle PL/SQL scripts called *Extreme_News_Algorithm*, defined in Figure 5.8. Parameter (*p_algo*) allows the user to call specific macro. Any number of extreme new extraction algorithms can be defined.

```
Procedure Extreme_News_Algorithm ( p_threshold in number default 0.75,
                                   p_algo   in varchar2 default 'ESE_VOL'
                                  )
```

Figure 5.8 Extreme_News_Algorithm algorithm definition

The prototype implements five algorithms, four extreme ranking algorithms to extract extreme news sentiment records, which are summarised in Table 5.7. An additional fifth algorithm also defined (referred to as ALL_NEWS) can be used for benchmarking against the other four algorithms. More detailed description and the pseudo code for these algorithms is shown in Appendix C.

Table 5.7 Prototype Extreme_News_Algorithm algorithms

□◣ ▲✕▛ ◡▛ □□▛ ◂ ?□□◜ ■ ▨✕ ● ▲ ● ◡□ ● ✔ ◡▛ □	□◣ ▲✕▛ ◡▛ □□▛ ◂ ?□□◜ ■ ▨✕ ● ▲ ● ◡□
ESE_T1	Naïvely inspects only negatively tagged news for a effect of the positively tagged news records, which negative news. This is a method widely used by capital companies.
ESE_T2	Computes the difference of the means between neg and positively tagged news records for each day, an to be extremely negative if the difference is grea variable
ESE_VOL	Uses two counters, one counter counts news item second counts the negative news items for e TRNA_SCORES table. Then computes differenc the difference is more than a threshold parameter, with highly negative news ratio
ESE_TOT	Investigates the role of the sentiment weight of th count of news items (as implemented in ESE_VO the difference between the sum of the sentiment items and the sum of the sentiment scores of difference is more than a certain threshold then th extreme news record
ALL_NEWS	This is a naïve algorithm retrieving all news Filtration_Function method (see subsection 5.2.2. algorithm as benchmark against other extreme alg

5.2.3 Implementing Impact Analysis Component

The IA component interactions with other components was shown in the sequence diagram in Figure 4.13. The user defines through the GUI the impact measure of the study, and triggers the IA component. Based on the impact measure parameter defined, the corresponding impact model will be invoked. The prototype implements four different impact models based on the impact measures described in subsection 4.2.3.4. Table 5.8 shows the different tools and software packages utilised for each impact model.

Table 5.8 Tools used in implementing the Impact models

Tool	**Impact Model**			
	Daily MCAAR	Intraday MCAAR	Intraday LBM	Intraday Price Jumps
Eventus	✓			
R Statistical Software		✓	✓	✓
PL/SQL Scripts	✓	✓	✓	✓
Target Entity	IMPACT_RESULTS	IMPACT_RESULTS	IMPACT_RESULTS	IMPACT_RESULTS

These impact models are explained as follows:

5.2.3.1 Implementing the Daily MCAAR impact model

The implementation of this impact model used Eventus software to compute the Daily Mean Cumulative Average Abnormal Returns (Daily MCAAR) figures (Rabhi et al., 2012; Yao and Rabhi, 2015). Eventus software is a tool that is based on SAS statistical analysis programming language and widely used to conduct event studies by finance researchers. The Impact Analysis (IA) component calls Eventus with the appropriate parameters to compute abnormal returns around an event date, performs statistical significance tests using the Daily MCAAR figures and produces a text file summary report of the impact study results.

5.2.3.2 Implementing the Intraday MCAAR impact model

The Intraday Mean Cumulative Average Abnormal Return (MCAAR) impact model (see subsection 4.2.3.4) has been implemented as per the following steps:

1- MERGE_TSLOT_AR database object reads the parameters defined in IMPACT_STUDIES_LOG database object and calculate the expected returns using the following equation:

$$Expected_{Returns} = Intercept + (slope \times Index_{\text{Returns}}) \quad (5.1)$$

Where the *Slope* and *Intercept* are coefficients variables used for calculating the intraday abnormal returns. These are provided by Oracle database stored functions *REGR_SLOPE* and *REGR_INTERCEPT.*

2- MERGE_TSLOT_AR database object uses the expected returns ($Expected_{Returns}$) to calculate the abnormal returns as per the following equation:

$$Abnormal_{Returns} = (Actual_{Returns} - Expected_{Returns}) \quad (5.2)$$

3- A PL/SQL script is used to loop through MERGE_TSLOT_AR database object and calculate the Intraday Mean Cumulative Average Abnormal Returns figures. This step accumulates per each time slot (5 minutes each) the Mean Cumulative Average Abnormal Returns (MCAAR) and saves the MCAAR results into the IMPACT_RESULTS database object.

4- An R software script is used to read the IMPACT_RESULTS database object and calculate the statistical significance of the MCAAR figures using Parametric One Sample T Test (Frost, 2016).

5- A PL/SQL script is used to update IMPACT_RESULTS database object with the test results.

5.2.3.3 Implementing the Intraday Liquidity Based Model (Intraday LBM)

This model is implemented using (Gomber et al., 2015) method (see subsection 4.2.3.4) as per the following steps:

1- A PL/SQL script is used to reads MERGE_MARKETDEPTH_TIMESLOTS database object passing the parameters which have been defined in the IMPACT_STUDIES_LOG database object through the GUI. This database object calculates the *median* of XLM figures, then saves the results to IMPACT_RESULTS database object.

2- An R software script is written to read the IMPACT_RESULTS database object and calculate the statistical significance of the Intraday LBM figures using Nonparametric Wilcox one sample test (Statistics Solutions, 2017).
3- A PL/SQL script is used to update IMPACT_RESULTS database object with the test results.

5.2.3.4 Implementing the Intraday Price Jumps impact model

This model implemented (Lee & Mykland, 2007) method to detect intraday price jumps (see subsection 4.2.3.4) as per the following steps:

1- A PL/SQL script is used to reads MERGE_PRICEJUMPS_TIMESLOTS database object passing to it the parameters which have been defined in the IMPACT_STUDIES_LOG database object through the GUI. This database object calculates the *median* of *JUMPSTATISTIC* figures, then saves the results to IMPACT_RESULTS database object.
2- An R software script is written to read the IMPACT_RESULTS database object and calculate the statistical significance of the Intraday Price Jumps figures using Nonparametric Wilcox one sample test (Statistics Solutions, 2017).
3- A PL/SQL script is used to update IMPACT_RESULTS database object with the test results.

5.3 Implementing the GUI layer

The GUI layer comprises interfaces, one interface for each use case (see section 4.4). It has been developed using Java Swing (ZetCode, 2017). These GUIs are explained as follows:

5.3.1 Define the FC parameters

The GUI shown in Appendix B (Figure B. 1 and Figure B. 2) illustrates the steps to initiate and define Financial Context parameters and save the study, which will generate a unique study number. The user will define the Entity (E), Benchmark (B), Entity Variable (EV) and Benchmark Variable (BV) parameters. Once the “Save and Continue” button is pressed, it would invoke the DMM component to log the parameters to the IMPACT_STUDIES_LOG database object as explained by the use case (see subsection 4.4.2).

5.3.2 Define the sentiment extraction parameters

The GUI shown in Appendix B (Figure B. 3) enables the user to define the sentiment extraction parameters for the saved study. The GUI implements number of filtration functions. Each function defines a number of filtration attributes, upon selecting the filtration function, the corresponding attributes gets populated. The user can then edit the filtration attributes 'values. The GUI calls the DMM component to log the parameters to the IMPACT_STUDIES_LOG database object. The DMM component will invoke the implemented SP component to apply the sentiment extraction algorithms (see subsection 5.2.2).

5.3.3 Conduct Impact Analysis use case

The GUI shown in Appendix B (Figure B. 4) enables the user to define the impact measure parameters and trigger the impact study. The user defines the type of the impact measure, then based on the impact measure type the user defines other related parameters, such as the estimation window and the event window. in the case of Daily and Intraday (MCAAR). The user then clicks the "Save IM parameters" button to invoke the DMM component (see subsection 4.4.4). This component will log the IM parameters to the IMPACT_STUDIES_LOG database object. Then the user clicks the "Conduct Impact Analysis Study" button to trigger the Impact Analysis (IA) component. The IA component will calculate the impact using the impact analysis API selected and save the results to IMPACT_RESULTS database object.

5.4 Limitations

Designing and implementing the prototype discussed in this chapter, was surrounded with challenges. We could summarize the major limitations we have with the current implementation of the prototype as:

- The implemented prototype is based on one sentiment dataset imported from Thomson Reuters. Future releases of the prototype should be flexible to accommodate more than one sentiment dataset. For example, it would be very useful to extend the prototype with datasets from Raven Pack or Bloomberg.

- The prototype implemented four extreme sentiment extraction algorithms but an IT expert is still required to implement new Extreme Sentiment Extraction (ESE) algorithms.. Further work is needed to find a way to eliminate the need for programming skills in order to be able to introduce new algorithms.
- The prototype implemented four impact models, however, an expert in data modelling is still needed to define new impact models, such as for instance volatility models, because this requires extensive changes to existing relationships. Further work is needed to find a new way to allow the analyst to define new impact models without the help of a data modeling expert.

5.5 Conclusion

The goal of implementing a prototype of the News Sentiment Impact Analysis (NSIA) framework was to demonstrate that it is possible to build a practical and feasible solution to automate many of the processes involved in sentiment data impact testing. The prototype utilises different technologies to integrate heterogenous tools and perform analysis computations. Users can follow well-defined use cases to conduct their analysis studies. In addition, the prototype sourced both high frequency market data and sentiment data from Thomson Reuters, which made it possible to design realistic impact evaluation experiments.

The implemented prototype conforms with the third stage (implementation stage) of the research process described in section 3.4. Chapter 6 will now describe real-life case studies which use the prototype and which are part of the fourth stage (evaluation stage).

6 NSIA FRAMEWORK EVALUATION

In accordance with the research process described in section 3.4, the NSIA framework will be evaluated using three case studies. Section 6.1 gives an overview of how these case studies relate to the research objectives. The first case study is detailed in section 6.2, while the second case study is described in section 6.3. The third and final case study is detailed in section 6.4. The overall results are then discussed in section 6.5, before the chapter is concluded in section 6.6.

6.1 NSIA framework evaluation

As discussed in chapter 3, the motivations behind designing the NSIA framework are as follows:

- The novel data model simplifies the identification and representation of the parameters used in sentiment-driven impact analysis studies, and addresses the *flexibility* issue by providing users with a way to set different contexts for impact analysis. Each context is uniquely identified by a set of parameters divided into Financial Context (FC), Sentiment Extraction (SN) and Impact Measure (IM).
- The use cases make the job of conducting experiments repeatable for the users and address the *reproducibility* issue by allowing users to repeat impact analysis studies in a consistent way.
- The software architecture facilitates the automation of impact analysis studies and allows the *reuse* and *interoperability* of existing software components, libraries, and packages when conducting impact analysis studies.

For the evaluation of the framework, three case studies have been selected as follows:

- First case study uses negative news daily impact on two companies to conduct an initial evaluation of the proposed framework.
- Second case study devises more complex scenarios, where multiple financial contexts are utilized.
- Third case study evaluate the framework utilizing different impact measures (intraday measures).

In Table 6.1 the case studies are illustrated according to which research objective is being addressed. The rest of this chapter now describes these case studies in more detail.

Table 6.1 Relating case studies and research objectives

Evaluation Criteria	**Case Study1 Negative News Daily Impact on two companies**	**Case Study2 Negative News Daily Impact in multiple contexts**	**Case Study3 Negative News Intraday Impact**
Flexibility/ Extensibility	Ability to define two FC and SN parameters	Ability to define multiple FC and SN parameters	Ability to define multiple IM parameters
Reproducibility/ Consistency	Ability to conduct a single daily impact analysis	Ability to carry out daily impact analysis according to different FC and SN parameters	Ability to carry out intraday impact analysis using different IM parameters
Automation/ Interoperability	Ability to import data and invoke a daily impact analysis software	Ability to filter and process data according to different FC and SN parameters	Ability to invoke different software packages according to different IM parameters.

6.2 Case study 1: Negative News Daily Impact on two companies

The goal of this case study is to conduct a preliminary validation of the NSIA framework, by assessing its readiness and functionality. The case study is described from a financial analyst or finance researcher's perspective. The objective of the analyst is to assess the efficacy of negative news (selected from Thomson Reuters News Analytics dataset) in measuring the impact of daily closing prices for two companies (BHP Billiton and Qantas). The case study's objective is to demonstrate that the NSIA framework provides the analyst with the *flexibility* to define two financial contexts (two companies) and utilize different filtration and sentiment extraction algorithms.

6.2.1 Defining the CPD model parameters

First, we must identify all CPD parameters, which are part of this case study. According to the NSIA framework, these parameters are divided between FC, SN and IM parameters (see subsection 4.2.3). Table 6.2 shows the FC parameters that are part of this case study.

Table 6.2 Defining Financial Context (FC) parameters

Financial Context (FC)	RIC Value	Description
Entity (E)	QAN.AX or BHP.AX	Qantas airlines, BHP Billiton – Mining giant
Entity Variable (EV)	Daily Closing price	Daily closing price of company
Benchmark (B)	. AORDA	All ordinaries market index which lists top 500 Australian listed companies is used as a benchmark (Australian Stock Exchange, 2014)
Benchmark Variable (BV)	Daily Closing price	Daily closing value of index
Study Period (P)	(1/01/2011, 31/12/2011)	The period during which the experiment takes place. For the purpose of this case study we assume one-year worth of news data is considered large enough to work with

In this case study, we wish to experiment with three different Filtration Functions (FA) and two Extreme Sentiment Extraction (ESE) algorithms. The ESE_T1 algorithm naively considers news as extreme if the negative sentiment score of news items for a day exceeds a certain threshold. The ESE_T2 algorithm only considers news items where the difference between the means of negative and positive tagged news items exceed a certain threshold, for more details see subsection 5.2.2. Accordingly, the SN parameters that are part of this case study are illustrated in Table 6.3.

Table 6.3 Defining the SN parameters

Sentiment Extraction (SN) parameters	SN Parameter value	Description (see subsection 5.2.2)
Filtration Function (FA)	Filtration Function 1	Select news where RIC = QAN.AX or BHP.AX, R=1, SRD= [01/01/2011, 31/12/2011], NT = ''
	Filtration Function 2	Select news where RIC = QAN.AX or BHP.AX, R=1, SRD= [01/01/2011, 31/12/2011], NT = 'O'
	Filtration Function 3	Select news where RIC = QAN.AX or BHP.AX, R=1, SRD= [01/01/2011, 31/12/2011], NT = 'JOB'
Extreme Sentiment Extraction (ESE) algorithm	ESE_T1	Selects news for a day where sentiment class is negative and sentiment score greater than a threshold value.

	ESE_T2	Selects news for a day where the difference between the mean of negative and positive news for a day passes certain threshold value.

Finally, the analyst needs to define the IM parameters. Since this case study is concerned with analysing daily impact, the Daily Mean Cumulative Average Abnormal Returns (Daily MCAAR) (see subsection 4.2.3.4) is selected as the impact measure. This measure requires defining two additional parameters: the event window and the estimation window as shown in Table 6.4.

Table 6.4 Defining the IM parameters

Impact Measure parameter	**Parameter Value**	**Description**
IM parameter	Daily MCAAR	Daily Mean Cumulative Average Abnormal Returns
Event Window	(0,0)	0 denotes the news release date
Estimation Window	(-30 Days, +30 Days)	Denotes the period of 30 days before the news release data and 30 days after the news release date
Time series frequency	Daily	Determine the frequency of the generated timeseries data

6.2.2 Performing the use cases

Having identified the parameters, the analyst is able to perform the three use cases that are supported by the prototype (see section 4.4) as follows:

- In the first two use cases, the analyst uses the provided GUIs (see subsections 5.3.1 and 5.3.2) to define the FC and SN parameters respectively. This results in the identification of the datasets required for an impact study. As there are some variations in some of the parameters, this results in eight distinct impact studies as shown in Table 6.5. For simplicity, we are referring to the subset of news obtained after applying the filtering function as FSP and the subset of extreme news as ESP. Table 6.5 shows the size of these datasets for each impact study.
- In the Conduct Impact Analysis use case, the analyst selects the impact measure parameters using the provided GUI (see subsection 5.3.3) and launches the impact analysis process. As already described in subsection 5.2.3.1, this prototype will invoke the Eventus software (Cowan Research, 2016), to compute the Daily Mean Cumulative Average Abnormal Returns (Daily MCAAR) figures. The results file (produced by Eventus) is then presented to the user.

Table 6.5 Different CPD parameters used in case study 1

Study No.	Entity (E)	Sentiment Parameters		Dataset size	
		Filter_News(FA)	ESE_Tn Algorithm	|FSP[1]|	|ESP[2]|
1	QAN.AX	Filtration Function 1	ESE_T1	203	83
2	QAN.AX	Filtration Function 2	ESE_T1	14	7
3	QAN.AX	Filtration Function 3	ESE_T1	56	20
4	BHP.AX	Filtration Function 1	ESE_T1	322	132
5	BHP.AX	Filtration Function 2	ESE_T1	31	7
6	BHP.AX	Filtration Function 3	ESE_T1	26	14
7	QAN.AX	Filtration Function 1	ESE_T2	203	19
8	BHP.AX	Filtration Function 1	ESE_T2	322	12

FSP^1 stands for FILTRATION_SN_PARAM table, ESP^2 stands for EXTREME_SN_PARAM table.

6.2.3 Results discussion

The eight impact studies conducted produced the results displayed in Table 6.6. The Sentiment Score Statistics column gives the number of extreme news as well as the number of distinct days, as it is possible to have multiple extreme news in the same day. Since we are using the daily impact evaluation technique, the impact is measured per day, the exact timing of the extreme news events not being taken into account. Daily MCAAR column shows the daily abnormal returns of entity (E) relative to benchmark (B). The Precision Weighted (CAAR), Patell Z and Generalized Sign Z columns highlight the statistical significance of the returns. The symbols $, *, **, and *** denote the statistical significance at the 0.10, 0.05, 0.01 and 0.001 levels, respectively, using a generic one-tail test for non-zero MCAAR successfully.

Table 6.6 Impact Studies Results

Study No.	Entity (E)	Sentiment Scores Statistics		Impact Measure Results				
		|ESP|	|D|	Event window	Daily Mean Cumulative Average Abnormal Return (MCAAR)	Precision Weighted (CAAR)	Patell Z	Generalized Sign Z
1	QAN.AX	83	48	(0,0)	+2.29%	2.22%	+0.842	+0.706
2	QAN.AX	7	7	(0,0)	+4.52%	4.18%	+2.166*	1.599$
3	QAN.AX	20	12	(0,0)	+0.77%	+1.25%	+0.488	-0.332
4	BHP.AX	132	98	(0,0)	-0.21%	-0.20%	-1.391$	-2.368**
5	BHP.AX	7	7	(0,0)	+0.14%	0.27%	0.626	-1.112
6	BHP.AX	14	10	(0,0)	-0.28%	-0.29%	-0.741	-1.829*
7	QAN.AX	19	4	(0,0)	-0.12%	-0.26%	-0.476	+ 0.423
8	BHP.AX	12	5	(0,0)	-0.51%	-0.51%	-1.699*	-2.020*

Based on the results, the analyst highlighted the following insights:

- Impact studies 1 to 6 which used a naïve approach for selecting extreme news i.e. simply looking at values of sentiment scores beyond a threshold (ESE_T1 algorithm) show that measuring the impact using Daily MCAAR is not meaningful, because it doesn't take into account positive news occurring on the same day as negative news. This results in having positive MCAAR in four studies (1, 2, 3, 5). For instance, six news stories with high positive sentiment scores were released about BHP responding to the labour strikes news released on the same days. Similarly, Qantas issued four positive news items about negotiating and resolving issues with workers' unions in relation to employees' benefits. These positive news stories were a counter response by both Qantas and BHP to negative news, which prevented the returns on those days from falling below beyond the benchmark.
- Impact studies 7 and 8 show the results of using a better technique for selecting extreme news (ESE_T2 algorithm), which considers positive news in the same day as negative news (see subsection 5.2.2.2). This algorithm resulted in negative (MCAARs) as expected, but after reading the negative news stories, it was found that they were not highly relevant to BHP and Qantas. Therefore, the expected decline in returns wasn't very significant.

- Daily MCAAR figures are affected by the filtration function (FA parameter), as we can observe in impact studies 2 and 3 for Qantas that news related to strikes (Filtration Function 3) had more impact than news about oil (Filtration Function 2). For QANTAS, we can see that using no topic filter (Filtration Function 1) in study 1 gives better results than filtering using Filtration Function 3 in study number 3 and worse than filtering using Filtration Function 2 in study number 2. The same applies when comparing studies 4, 5 and 6 for BHP.

6.2.4 Discussion

The motivation behind this case study was to evaluate *Flexibility* of the proposed framework in setting different parameters for conducting an impact study. The prototype enabled the analyst to define eight simple financial contexts (related to two companies), and employ different sentiment filtration functions, as well as extreme news selection techniques and conduct eight impact analysis studies. The analyst was also able to derive simple insights from the results, which show the limitation of using daily impact analysis for a dataset in which multiple news with different sentiment scores may happen on the same day.

The *interoperability* of the NSIA framework has been demonstrated as it was possible to integrate a real-life news sentiment dataset with trading data as well as conduct daily impact analysis using the Eventus software (which runs on a SAS platform). The prototype's support for automation was also demonstrated as it was possible to generate impact evaluation results from the CPD parameters defined by the analyst via the GUIs. The impact studies' results are in in text format, which then have to be read and interpreted by the analyst.

6.3 Case Study2: Negative news daily impact in multiple contexts

This case study is designed to demonstrates the *flexibility* of the NSIA framework when conducting more complex impact studies. It evaluates the impact of negative sentiment news in multiple financial contexts, and extending the sentiment filtration capabilities to employ more advanced news sentiment filtration algorithms.

6.3.1 Defining the CPD model parameters

In this case study, we wish to study the impact of news on twenty-six companies in four different markets (Australia, Germany, Canada, USA). In the Australian market, the companies selected will be the constituents of ATLI index. For the German market, the companies selected are the common constituents of GDAXI and CXKNX indices. In the Canadian market, the companies are the common constituents of GTSX and SPTSE indices. In the USA market, the companies selected are the common constituents of DJI and HWI indices. Table 6.7 shows the four groups of companies selected as part of this case study.

Table 6.7 Financial context entities

Entity E	**Common constituents in each context**
{ATLI}	NCM.AX➔ Newcrest Mining Limited QBE.AX➔ QBE Insurance Group Limited ANZ.AX➔ The Australia and New Zealand Banking Group Limited AMP.AX➔ AMP WES.AX➔ Wesfarmers Limited SUN.AX➔ Suncorp Group Limited CSL.AX➔CSL Limited ORG.AX➔Origin is an Australian energy company TLS.AX➔ Telstra Corporation Limited BXB.AX➔ Brambles Limited (holding company of the Brambles Group) WOW.AX➔ Woolworths Limited FGL.AX➔ Foster's Group Ltd NAB.AX➔ National Australia Bank WPL.AX➔ Woodside Petroleum Limited WRT.AX➔ Westfield Retail Trust WBC.AX➔ Westpac Banking Corporation CBA.AX➔ The Commonwealth Bank of Australia MQG.AX➔Macquarie Group Limited WDC.AX➔WestfieldG Stapl
{GDAXI ∩ CXKNX}	TKAG.DE➔ Thyssenkrupp AG MANG.DE➔ MAN SE (German manufacturer vehicles and engines) SIEGn.DE➔Siemens AG

{GTSX ∩ SPTSE}	VRX.TO➔ Valeant Pharmaceuticals International, Inc. BV.TO➔ Biovail Corp (Pharmaceuticals)
{DJI ∩ HWI}	IBM.N➔ International Business Machines Corp HPQ.N➔ The Hewlett-Packard Company

We will also use two benchmarks for every group of companies. The period of study will be increased to two years. Table 6.8 shows the FC parameters defined for this case study.

Table 6.8 Defining Financial Context (FC) parameters

Financial Context (FC)	**RIC Value**	**Description**
Entity (E)	{ATLI}	Constituents of the ATLI index
	{GDAXI ∩ CXKNX}	Constituents that are listed on both GDAXI and CXKNX indices
	{GTSX ∩ SPTSE}	Constituents that are listed on both GTSX and SPTSE indices
	{DJI ∩ HWI}	Constituents that are listed on both DJI and HWI indices
Entity Variable (EV)	Daily Closing price	Daily closing price of company
Benchmark (B)	ATLI	Australia's ASX top 20 leaders
	AORD	Australia's ASX All Ordinaries Index
	CXKNX	Germany's Industrial Index
	GDAXI	Germany's DAX Index
	GTSX	Canada's healthcare index
	SPTSE	S&P Toronto Stock Exchange
	DJI	Dow Jones Industrial Index in the USA
	HWI	NYSE Arca Computer Hardware Index in the USA
Benchmark Variable (BV)	Daily Closing price	Daily closing value of index
Study Period (P)	(1/01/2010, 31/12/2011)	The period during which the experiment takes place. For the purpose of this case study we assume two years worth of news data is considered large enough to work with

In this case study, we wish to experiment with one new Filtration Function (FA) and three new Extreme Sentiment Extraction (ESE) algorithms. These sentiment extraction algorithms are already predefined in the prototype (see subsection 5.2.2), they are called ESE_VOL, ESE_TOT and ALL_NEWS. The first two ESE algorithms (ESE_VOL, ESE_TOT) will use the following two parameters:

- Pr: threshold parameter to define the ratio between negative and positive news for a day. In this case study, we will be using the values 0.2 and 0.3 (the higher the threshold value the *more* news items are detected by the ESE algorithm).
- Ps: threshold parameter to test how far a news item's sentiment score value should be close or far from the mean of sentiment score in a given day. In this case study, we will be using the values 0.5 and zero (the higher the threshold value the *less* news items are detected by the ESE algorithm).

Accordingly, the SN parameters that are part of this case study are illustrated in Table 6.9.

Table 6.9 Defining the SN parameters

Sentiment Extraction (SN) parameters	**SN Parameter value**	**Description (see subsection 5.2.2)**
Filtration Function (FA)	Filtration Function 4	RIC= {ATLI ∩ AORD} or {GDAXI ∩ CXKNX} or {GTSX ∩ SPTSE} or {DJI ∩ HWI}, R=1, NC=1, NV=0, SRD= [1/01/20110,31/12/2011]
Extreme Sentiment Extraction (ESE) algorithm	ALL_NEWS	Naively selects all news for a day related to Entity E
	ESE_VOL (Pr, Ps, TRNA_SCORES)	Selects news for a day where ratio between volume of negative news and positive news is less than a threshold value (Pr) and the (Ps) threshold value determines how far a news item's sentiment score value should be close or far from the mean of sentiment
	ESE_TOT (Pr, Ps, TRNA_SCORES)	Selects news for a day where ratio between total sentiment scores of negative news and positive news is less than the threshold values (Pr) and the (Ps) threshold value determines how far a news item's sentiment score value should be close or far from the mean of sentiment

Finally, this case study utilizes the same impact measure parameters that were defined in the first case study. It uses the Daily Mean Cumulative Average Abnormal Returns (Daily MCAAR) as the impact measure (see subsection 4.2.3.4) as shown in Table 6.10.

Table 6.10 Defining the IM parameters

Impact Measure parameter	**Parameter Value**	**Description**
IM parameter	Daily MCAAR	Daily Mean Cumulative Average Abnormal Returns

Event Window	(0,0)	0 denotes the news release date
Estimation Window	(-30 Days, +30 Days)	Denotes the period of 30 days before the news release data and 30 days after the news release date
Time series frequency	Daily	Determine the frequency of the generated timeseries data

6.3.2 Performing the use cases

For this case study, the analyst is able to perform the three use cases (see section 4.4) as follows:

- In the first two use cases, the analyst uses the same GUIs as in the first case study to define the FC and SN parameters respectively. This results in twenty-four distinct impact studies as shown in Table 6.11. Each FC was evaluated three times using the three Extreme Sentiment Extraction (ESE_VOL, ESE_TOT and ALL_NEWS) algorithms.
- Like in the first case study, the analyst launches the Conduct Impact Analysis use case by defining the impact measure parameters and launching the impact analysis process. Likewise, as described in subsection 5.2.3.1, this prototype will invoke the Eventus software, to compute the Daily Mean Cumulative Average Abnormal Returns (Daily MCAAR) figures. The results file is then presented to the user.

Table 6.11 Different CPD parameters used in case study 2

Study No.	**Entity (E)**	**Sentiment Parameters**		**Dataset size**	
		Filter_News(FA)	ESE_Tn Algorithm	$\|FSP^1\|$	$\|ESP^2\|$
1	WDC.AX, ANZ.AX, TLS.AX, WES.AX, WPL.AX, NCM.AX, CBA.AX	Filtration Function 4	ESE_VOL	1333	30
2	WES.AX, WBC.AX, WPL.AX, ANZ.AX, TLS.AX, ORG.AX, NCM.AX, QBE.AX, MQG.AX, WOW.AX	Filtration Function 4	ESE_TOT	1333	28
3	AMP.AX, ANZ.AX, BXB.AX, CBA.AX, CSL.AX, FGL.AX, MQG.AX, NAB.AX, NCM.AX, ORG.AX, QBE.AX, SUN.AX, TLS.AX, WBC.AX, WDC.AX, WES.AX, WOW.AX, WPL.AX, WRT.AX	Filtration Function 4	ALL_NEWS	1333	1333
4	WDC.AX, TLS.AX, WES.AX, ANZ.AX, WPL.AX, CBA.AX, NCM.AX	Filtration Function 4	ESE_VOL	1333	30

5	WES.AX, WBC.AX, WPL.AX, ANZ.AX, TLS.AX, ORG.AX, NCM.AX, QBE.AX, MQG.AX, WOW.AX	Filtration Function 4	ESE_TOT	1333	28
6	AMP.AX, ANZ.AX, BXB.AX, CBA.AX, CSL.AX, FGL.AX, MQG.AX, NAB.AX, NCM.AX, ORG.AX, QBE.AX, SUN.AX, TLS.AX, WBC.AX, WDC.AX, WES.AX, WOW.AX, WPL.AX, WRT.AX	Filtration Function 4	ALL_NEWS	1333	1333
7	MANG.DE, SIEGn.DE, TKAG.DE	Filtration Function 4	ESE_VOL	776	37
8	MANG.DE, SIEGn.DE, TKAG.DE	Filtration Function 4	ESE_TOT	776	34
9	MANG.DE, SIEGn.DE, TKAG.DE	Filtration Function 4	ALL_NEWS	776	776
10	MANG.DE, SIEGn.DE, TKAG.DE	Filtration Function 4	ESE_VOL	776	37
11	MANG.DE, SIEGn.DE, TKAG.DE	Filtration Function 4	ESE_TOT	776	34
12	MANG.DE, SIEGn.DE, TKAG.DE	Filtration Function 4	ALL_NEWS	776	776
13	BVF.TO	Filtration Function 4	ESE_VOL	119	2
14	BVF.TO, VRX.TO	Filtration Function 4	ESE_TOT	119	10
15	BVF.TO, VRX.TO	Filtration Function 4	ALL_NEWS	119	119
16	BVF.TO	Filtration Function 4	ESE_VOL	119	2
17	BVF.TO, VRX.TO	Filtration Function 4	ESE_TOT	119	10
18	BVF.TO, VRX.TO	Filtration Function 4	ALL_NEWS	119	119
19	IBM.N, HPQ.N	Filtration Function 4	ESE_VOL	1183	28
20	HPQ.N	Filtration Function 4	ESE_TOT	1183	15
21	IBM.N, HPQ.N	Filtration Function 4	ALL_NEWS	1183	1183
22	IBM.N, HPQ.N	Filtration Function 4	ESE_VOL	1183	28

23	HPQ.N	Filtration Function 4	ESE_TOT	1183	15
24	IBM.N, HPQ.N	Filtration Function 4	ALL_NEWS	1183	1183

FSP[1] stands for FILTRATION_SN_PARAM table, ESP[2] stands for EXTREME_SN_PARAM table.

6.3.3 Results discussion

The twenty-four impact studies conducted produced the results displayed in Table 6.12. Column |D| is the number of distinct event days determined by applying the ESE algorithm (ESE_VOL and ESE_TOT). As in the previous case study, the number of extreme news can be different from the number of distinct days, as it is possible to have multiple extreme news in the same day. As before, Daily MCAAR column shows the daily abnormal returns of entity (E) relative to benchmark (B). The Precision Weighted (CAAR), Patell Z and Generalized Sign Z columns highlight the statistical significance of the returns. The symbols $, *, **, and *** denote the statistical significance at the 0.10, 0.05, 0.01 and 0.001 levels, respectively, using a generic one-tail test for non-zero MCAAR successfully.

Table 6.12 Impact Studies Results

Study No.	**Country**	**Entity (E)**	**Sentiment Scores Statistics**		**Impact Measure Results**				
			|ESP|	|D|	Event window	Daily MCAAR	Precision Weighted (CAAR)	Patell Z	Generalized Sign Z
1	Australia	{ATLI ∩ AORD}	30	5	(0,0)	-0.18%	-0.03%	-0.078	-0.216
2	Australia	{ATLI ∩ AORD}	28	12	(0,0)	-0.84%	-1.05%	-4.601***	-1.870*
3	Australia	{ATLI ∩ AORD}	1333	406	(0,0)	-0.05%	-0.06%	-1.571$	-0.007
4	Australia	{ATLI ∩ AORD}	30	5	(0,0)	-0.25%	-0.12%	-0.381	0.265
5	Australia	{ATLI ∩ AORD}	28	12	(0,0)	-0.82%	-1.05%	-4.496***	-1.383$
6	Australia	{ATLI ∩ AORD}	1333	406	(0,0)	-0.04%	-0.06%	-1.635$	-0.284

7	Germany	{GDAXI ∩ CXKNX}	37	12	(0,0)	-0.39%	0.01%	0.027	0.200
8	Germany	{GDAXI ∩ CXKNX}	34	9	(0,0)	0.07%	0.13%	0.619	-0.661
9	Germany	{GDAXI ∩ CXKNX}	776	298	(0,0)	0.16%	0.15%	2.792**	2.016*
10	Germany	{GDAXI ∩ CXKNX}	37	6	(0,0)	-0.31%	0.00%	-0.020	0.926
11	Germany	{GDAXI ∩ CXKNX}	34	9	(0,0)	0.28%	0.36%	1.066	0.622
12	Germany	{GDAXI ∩ CXKNX}	776	298	(0,0)	0.11%	0.08%	2.254*	1.453$
13	Canada	{GTSX ∩ SPTSE}	2	1	(0,0)	-0.84%	-0.84%	-0.711	-1.022
14	Canada	{GTSX ∩ SPTSE}	10	7	(0,0)	-0.22%	-0.15%	-0.597	-0.473
15	Canada	{GTSX ∩ SPTSE}	119	78	(0,0)	0.02%	0.05%	0.223	0.315
16	Canada	{GTSX ∩ SPTSE}	2	1	(0,0)	-1.23%	-1.23%	-0.753	-0.874
17	Canada	{GTSX ∩ SPTSE}	10	7	(0,0)	-0.07%	-0.10%	-0.114	1.353$
18	Canada	{GTSX ∩ SPTSE}	119	78	(0,0)	-0.02%	-0.01%	-0.502	-0.291
19	USA	{DJI ∩ HWI}	28	10	(0,0)	-0.10%	-0.29%	-1.289$	0.575
20	USA	{DJI ∩ HWI}	15	6	(0,0)	0.69%	0.22%	0.507	-0.265
21	USA	{DJI ∩ HWI}	1183	406	(0,0)	-0.02%	-0.03%	-0.991	1.833*
22	USA	{DJI ∩ HWI}	28	10	(0,0)	0.32%	0.30%	1.140	2.953**
23	USA	{DJI ∩ HWI}	15	6	(0,0)	0.95%	0.76%	1.649*	1.482$

24	USA	{DJI ∩ HWI}	1183	406	(0,0)	-0.03%	-0.04%	-1.091	0.501

Figure 6.1 shows the resulted aggregated Daily MCAAR figures according to which ESE algorithm is used. Impact studies utilizing ESE_VOL algorithm (studies 1,4,7,10,13,16,19,22) resulted in 3% cumulative drop. Drilling down by country, the results illustrated in Figure 6.2, show that the ESE_TOT algorithm performs better in Australia (studies 2 and 5, shows a negative MCAAR of 1.66%), whereas ESE_VOL algorithm performs better in the other countries. Figure 6.2 also shows that the Australian and Canadian markets were the best responsive to the sentiment dataset, as compared to the US and German markets. The results demonstrate how varying the ESE algorithms have produced different impact results. The benchmark algorithm ALL_NEWS (studies 3,6,9,12,15,18,21,24), which naively selects all news records regardless of their sentiment orientation, shows that stock returns figures do not react to this algorithm, which what we expected.

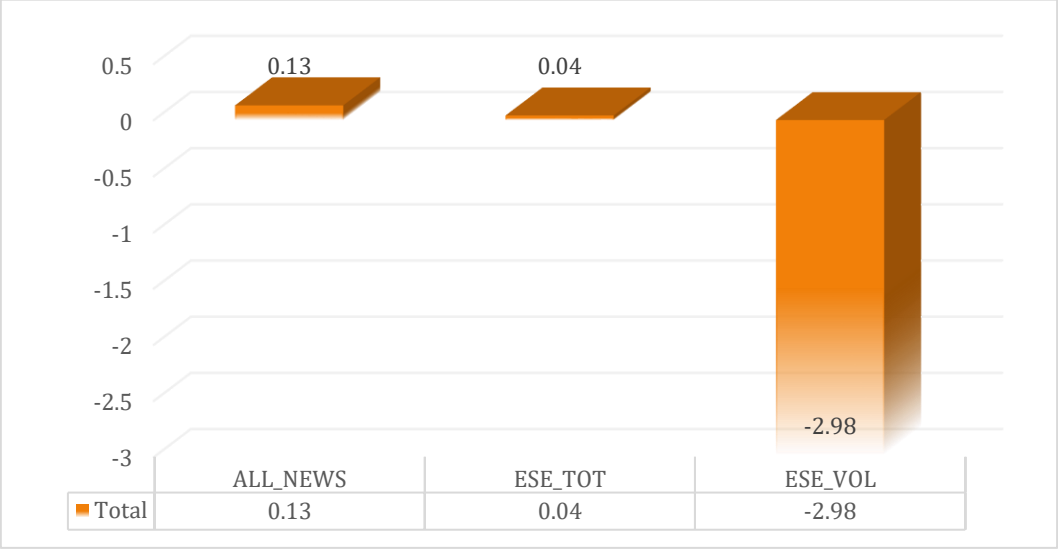


Figure 6.1 Sum of all MCARs for the 24 experiments by ESE Algorithms

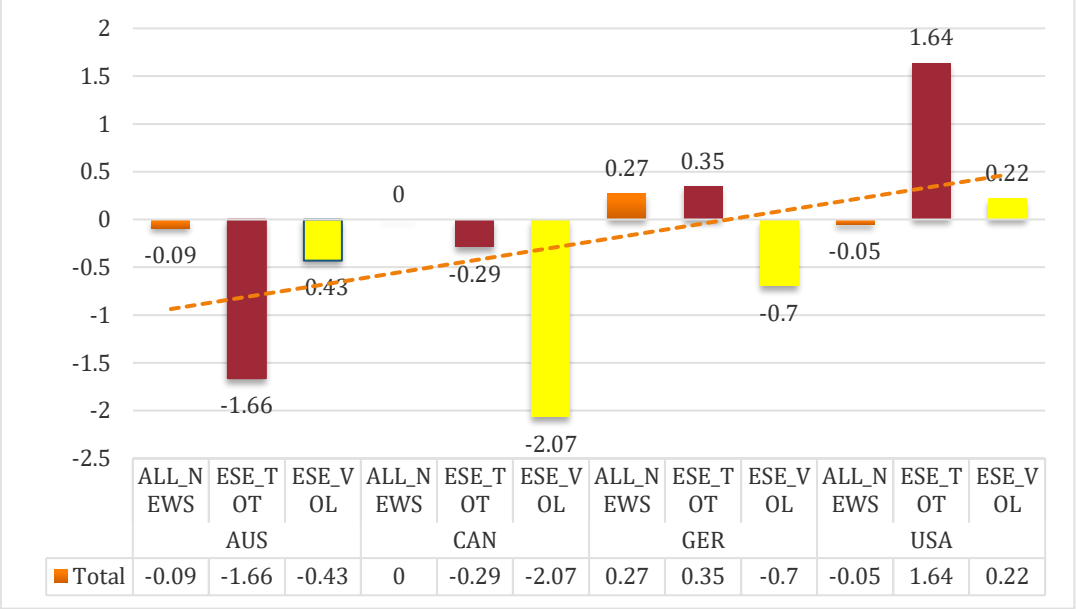


Figure 6.2 Impact grouped by country and Extreme Sentiment Extraction Algorithms

6.3.4 Discussion

This case study has tested most of the criteria used in the first case study. In addition, it demonstrated the ability to deal with a large number of contexts (twenty-four evaluation studies were conducted). From an analyst's perspective, these twenty-four studies demonstrated the capabilities of the proposed NSIA framework. The prototype has enabled selection of different groups of companies as well as extreme sentiment extraction algorithms, which gave the analyst important insights on the quality of these sentiment scores in different contexts.

6.4 Case Study3: Negative news intraday impact

6.4.1 Case study scenario

This case study constitutes the last iteration of NSIA framework's validation. Both of the previous case studies have used Daily Mean Cumulative Average Abnormal Returns (Daily MCAAR) as the impact measure whereas this case study is designed to demonstrate the ability of the NSIA framework to support different intraday impact measures. In this case study three impact models are used, namely intraday liquidity, intraday price jumps and intraday abnormal returns.

6.4.2 Defining the CPD model parameters

This case study reuses some of the financial context parameters defined in the second case study (see subsection 6.3.1). Firstly, Table 6.13 shows the financial contexts parameters used in this case study, which have six variations.

Table 6.13 Defining Financial Context (FC) parameters

Financial Context (FC)	**RIC Value**	**Description**
Entity (E)	{ATLI}	Constituents of the ATLI index in Australia
	{GDAXI∩ CXKNX}	Constituents that are listed on both GTSX and CXKNX indices in Germany
	{GTSX ∩ SPTSE}	Constituents that are listed on both GTSX and SPTSE indices in Canada
	{DJI ∩ HWI}	Constituents that are listed on both DJI and HWI indices in USA
Entity Variable (EV)	Trade price	Last trade price within 5 minute interval
Benchmark (B)	ATLI	Australia's ASX top 20 leaders
	AORD	Australia's ASX All Ordinaries Index
	GDAXI	Germany's DAX Index
	SPTSE	S&P Toronto Stock Exchange
	DJI	Dow Jones Industrial Average Index in the USA
	HWI	NYSE Arca Computer Hardware Index in the USA
Benchmark Variable (BV)	Index value	Last index value within a 5 minute interval
Study Period (P)	(1/01/2010, 31/12/2011)	The period during which the experiment takes place. For the purpose of this case study we assume two years worth of news data is considered large enough to work with

The case study reuses the same sentiment extraction parameters that were used in case study 2 (see Table 6.14), which has used two sentiment extraction algorithms and one filtration function.

Table 6.14 Defining the SN parameters

Sentiment Extraction (SN) parameters	SN Parameter value	Description (see subsection 5.2.2)
Filtration Function (FA)	Filtration Function 4	RIC= {ATLI ∩ AORD} or {GDAXI ∩ CXKNX} or {GTSX ∩ SPTSE} or {DJI ∩ HWI}, R=1, NC=1, NV=0, SRD= [1/01/20110,31/12/2011]
Extreme Sentiment Extraction (ESE) algorithm	ESE_VOL (Pr, Ps, TRNA_SCORES)	Selects news for a day where ratio between volume of negative news and positive news is less than a threshold value.
	ESE_TOT (Pr, Ps, TRNA_SCORES)	Selects news for a day where ratio between total sentiment scores of negative news and positive news is less than a threshold value.

Next, the case study uses three different *intraday* impact models, which are Liquidity Based Model (LBM), price jumps statistics and Mean Cumulative Average Abnormal Returns (MCAAR) (see subsection 5.2.3). The parameters used by these different models are shown Table 6.15.

Table 6.15 Defining the IM parameters

Impact Measure parameter	Impact Measure Type	Parameter name	Parameter Value	Description
IM parameter	Intraday MCAAR	Estimation Window	(-17 Days, -3 Days)	Denotes the period of 14 days before the news release date. This window is used to calculate the expected returns as clarified in the implementation of this model (see subsection 5.2.3.2)
	Intraday LBM	XLM	50000	The value of shares for each roundtrip transaction
	Intraday Price Jumps	Window Size	270	a positive integer which controls the window size of the price jump algorithm. The value 270 is recommended when using five-minute timeseries data (Lee & Mykland, 2007).
		Parameter C	0.7979	a positive or negative value that control the behavior of the price jump algorithm. The recommended parameter value is 0.7979 is used (Lee & Mykland, 2007)
		Threshold Value	4.6	a positive or negative threshold value that determine what is considered a price and what is not. The threshold value used is 4.6 (if the test statistic

				exceeds > 4.6, then this means there was price jump).
Event Window			{(-40 min, -10 min), (0 min, 30 min)}	0 denotes the time of news release
Time series frequency			Five-minute	Determine the frequency of the generated timeseries data

6.4.3 Performing the use cases

This case study executed the three use cases (see section 4.4) as follows:

- In the first two use cases, the analyst uses the same GUIs as in the first case study to define the FC and SN parameters respectively. Each FC parameters variation is evaluated two times this results in 12 distinct variations of financial context and sentiment parameters as shown in Table 6.16.
- The third use case Conduct Impact Analysis enabled the analyst to define three different impact measures and three event windows producing nine distinct variations. This resulted in twenty-eight distinct intraday impact studies (twelve intraday mean cumulative average abnormal returns, eight liquidity based measure and eight price jumps statistics). The results are then presented to the user in a text file.

Table 6.16 Different CPD parameters used in case study 3

Study No.	Entity (E)	Benchmark	Sentiment Parameters		Dataset size	
			Filter_News(FA)	ESE_Tn Algorithm	$\|FSP^1\|$	$\|ESP^2\|$
1	WDC.AX, ANZ.AX, TLS.AX, WES.AX, WPL.AX, NCM.AX, CBA.AX	ATLI	Filtration Function 4	ESE_VOL	1333	30
2	WES.AX, WBC.AX, WPL.AX, ANZ.AX, TLS.AX, ORG.AX, NCM.AX, QBE.AX, MQG.AX, WOW.AX	ATLI	Filtration Function 4	ESE_TOT	1333	28
3	WDC.AX, TLS.AX, WES.AX, ANZ.AX, WPL.AX, CBA.AX, NCM.AX	AORD	Filtration Function 4	ESE_VOL	1333	30

4	WES.AX, WBC.AX, WPL.AX, ANZ.AX, TLS.AX, ORG.AX, NCM.AX, QBE.AX, MQG.AX, WOW.AX	AORD	Filtration Function 4	ESE_TOT	1333	28
5	MANG.DE, SIEGn.DE, TKAG.DE	GDAXI	Filtration Function 4	ESE_VOL	776	37
6	MANG.DE, SIEGn.DE, TKAG.DE	GDAXI	Filtration Function 4	ESE_TOT	776	34
7	BVF.TO	SPTSE	Filtration Function 4	ESE_VOL	119	2
8	BVF.TO, VRX.TO	SPTSE	Filtration Function 4	ESE_TOT	119	10
9	IBM.N, HPQ.N	DJI	Filtration Function 4	ESE_VOL	1183	28
10	HPQ.N	DJI	Filtration Function 4	ESE_TOT	1183	15
11	IBM.N, HPQ.N	HWI	Filtration Function 4	ESE_VOL	1183	28
12	HPQ.N	HWI	Filtration Function 4	ESE_TOT	1183	15

FSP[1] stands for FILTRATION_SN_PARAM table, ESP[2] stands for EXTREME_SN_PARAM table.

6.4.4 Results discussion

6.4.4.1 Intraday MCAAR results

The results that correspond to selecting intraday MCAAR as the impact model are shown in Table 6.17. Column Country store the country of the market index and its constituents. Column Benchmark show the abbreviation for the market index name. Column |D| is the number of distinct event days decided by applying the ESE algorithm (ESE_VOL and ESE_TOT). As in the previous case study, the number of extreme news can be different from the number of distinct days, as it is possible to have multiple extreme news in the same day. Columns MCAAR (-40 min, -10 min) and MCAAR (0 min, 30 min) are two event windows that show the mean intraday MCAAR during these two windows. The statistical significance (see subsection 5.2.3) is tested using parametric Welch two sample t-test, which gives two figures (columns Welch two sample t- test and P value). The statistical significance of the results is highlighted with symbols $, *, **, and ***, which denote the statistical significance at the 0.10, 0.05, 0.01 and 0.001 levels, respectively.

Table 6.17 Intraday MCAAR impact results

Study no	Country	Benchmark	ESE Algo.	|ESP|	|D|	MCAAR (-40 min, -10 min)	MCAAR (0 min, 30 min)	Welch Two Sample T-Test	P Value
1	Australia	ATLI	ESE_VOL	30	5	3.65%	4.47%	-1.49	0.9067
2	Australia	ATLI	ESE_TOT	28	12	1.12%	2.18%	-3.74	0.9979
3	Australia	AORD	ESE_VOL	30	5	6.13%	-6.97%	4.51	0.00201**
4	Australia	AORD	ESE_TOT	28	12	1.23%	0.38%	6.79	0.00006***
5	Germany	GDAXI	ESE_VOL	37	6	-25.23%	-24.79%	-4.14	0.9992
6	Germany	GDAXI	ESE_TOT	34	9	3.66%	3.7%	-0.83	0.7857
7	Canada	SPTSE	ESE_VOL	2	1	2.99%	2.47%	3.04	0.01431**
8	Canada	SPTSE	ESE_TOT	10	7	2.33%	1.54%	8.89	0.00002***
9	USA	DJI	ESE_VOL	28	10	4.53%	3.29%	6.39	0.00008***
10	USA	DJI	ESE_TOT	15	6	9.63%	9.45%	2.13	0.06031$
11	USA	HWI	ESE_VOL	28	10	7.21%	5.86%	4.99	0.00101***
12	USA	HWI	ESE_TOT	15	6	8.87%	8.63%	1.77	0.07433$

In nine out of the twelve studies, intraday MCAAR dropped after releasing the news (window 0 to 30 minutes) compared with before releasing the news (window -40 to -10 minutes). In eight studies (study 3,4,7,8,9 and 11), the drop was significant at various levels. This provides enough evidence that there is strong correlation between the news filtered by ESE_VOL and ESE_TOT algorithms and the intraday (5 minutes interval) impact witnessed after the time of releasing the news.

Negative news volumes (ESE_VOL) algorithm generated better impact results than negative news weights (ESE_TOT) algorithm. This is evident in all contexts studied (across all countries). In addition, the number of days with significant high volumes is smaller than the number of days with lower volumes and higher sentiment weights (sentiment scores). This observation is true, except in the case of the USA markets, where the number of days with high volumes was higher than the number of days with extreme negative sentiment scores (ESE_TOT).

The results confirm that the ESE_VOL algorithm shows better results than the ESE_TOT algorithm, where out of the six impact studies using ESE_VOL, four generated statistically significant results (column P-Value). This result complies with findings in literature, that news volume is correlated with

higher/lower significant returns (Das and Chen, 2007). Impact studies using ESE_TOT algorithm generated significant results in also four out of the six studies (studies 4,8,10 and 12), with less significance. Out of four countries included in the case study, only Germany related studies didn't reflect any level of significance in their results (studies 5 and 6). This could be explained by the fact that English is not the formal language (news are released in German first then translated). This finding shows the importance of validating sentiment datasets across different financial contexts (countries, benchmarks), where assumptions made on the efficiency/deficiency of a sentiment dataset, could be confirmed/refuted, when testing against other financial contexts.

6.4.4.2 Intraday LBM results

The results that correspond to selecting intraday LBM as the impact model are shown in Table 6.18. Columns XLM measure (-40 min, -10 min) and XLM measure (0 min, 30 min) are two event windows that show the median of the intraday LBM figures during these two-time periods. The statistical significance (see subsection 5.2.3) is tested using non-parametric Wilcox two sample test, which gives two figures (columns Wilcox two sample test and P value). The statistical significance of the results is highlighted with symbols $, *, **, and ***, which denote the statistical significance at the 0.10, 0.05, 0.01 and 0.001 levels, respectively.

Table 6.18 Intraday LBM impact results

Study no	Country	SN	|ESP|	|D|	XLM measure (-40 min, -10 min)	XLM measure (0 min, 30 min)	Wilcox Two Sample Test	P Value
1	Australia	ESE_VOL	30	5	4.555	2.544	49	0.0013***
2	Australia	ESE_TOT	28	12	5.536	5.5095	31	0.4557
3	Germany	ESE_VOL	37	6	5.078	6.287	9	0.05303*
4	Germany	ESE_TOT	34	9	5.533	6.5635	7	0.02622*
5	Canada	ESE_VOL	2	1	16.914	16.9815	6	1
6	Canada	ESE_TOT	10	7	11.555	17.77	0	0.00058***
7	USA	ESE_VOL	28	10	4.249	6.431	12	0.124
8	USA	ESE_TOT	15	6	3.1965	7.544	0	0.01745**

The Intraday LBM results measure the cost of a round trip transaction with limit order of 50000 value (relevant to the country's currency). The hypothesis made is that extremely negative news would lead

to rise in the trading transactions cost, which leads to lower the liquidity. The results show that in almost all the studies seven out of the eight studies (see Table 6.18) show that XLM measures (column (0 min, 30 min)) have risen (except study number 1), which leads us to accept the hypothesis. This result is consistent with the existing literature, where liquidity decreases around negative news and increases around positive news (Riordan et al, 2013). The results in this table also correlate with the intraday MCAAR results (see Table 6.17). In one context (Germany) the results suggest that regardless of the extraction algorithm (ESE_VOL, ESE_TOT) used, the trading costs rise significantly (at the 5% level). In all other contexts (USA, Canada and Australia), the choice of ESE algorithm impact results differently, where the ESE_TOT algorithm's impact is significant in three out of the four studies (studies 4,6 and 8). The news volume (ESE_VOL) algorithm surprisingly doesn't seem to be have an impact (contrary to the intraday MCAAR results), where only in one study did the cost significantly spike reacting to significant rise in negative news volumes (study number 3). This finding is contradicted by significant drop in transactions costs in the Australian context (study 1), where trading costs plunge in response to a spike in negative news volumes (ESE_VOL algorithm). In general, we can say that news does have an impact on rising the trading transaction costs, especially when negative news are filtered by their weights (ESE_TOT) as opposed to their volumes (ESE_VOL). The results extend the body of literature (Riordan et al, 2013) on studying the impact of sentiment datasets on intraday liquidity, showing how intraday liquidity reacted to varying the financial context and/or employing various sentiment filtration techniques, impact the liquidity.

6.4.4.3 Intraday price jumps statistics results

The results that correspond to selecting intraday price jumps statistics as the impact model are shown in Table 6.19. Columns PJS (-40 min, -10 min) and PJS (0 min, 30 min) are two event windows that show the median of the intraday Price Jumps Statistics (PJS) during these two-time periods. The statistical significance (see subsection 5.2.3) is tested using non-parametric Wilcox two sample test, which gives two figures (columns Wilcox two sample test and P value). The statistical significance of the results is highlighted with symbols $, *, **, and ***, which denote the statistical significance at the 0.10, 0.05, 0.01 and 0.001 levels, respectively.

Table 6.19 Intraday price jumps statistics results

Study no	Country	SN	|ESP|	|D|	PJS (-40 min, -10 min)	PJS (0 min, 30 min)	Wilcox Two Sample Test	P Value
1	Australia	ESE_VOL	30	5	-17.8	-12.1	10	0.07301$
2	Australia	ESE_TOT	28	12	-13.22	-12.87	18	0.4557
3	Germany	ESE_VOL	37	6	-12.74	-13.1	33	0.3176
4	Germany	ESE_TOT	34	9	-13.35	-12.65	2	0.00233**
5	Canada	ESE_VOL	2	1	-8.39	-8.44	6	1
6	Canada	ESE_TOT	10	7	-13.57	-10.48	13	0.1594
7	USA	ESE_VOL	28	10	-13.91	-13.29	15	0.2593
8	USA	ESE_TOT	15	6	-15.22	-13.22	0	0.02857*

From these results, the following observations can be made:

- The price jumps statistics varied across the financial contexts (countries), where the highest impact is reflected on the German and USA markets (studies 4 and 8 respectively).
- The choice of sentiment filtration technique employed does have a varying effect on the results. The results confirm the ESE_TOT algorithm generated better results than the ESE_VOL algorithm. This result also agrees with the findings of the intraday LBM (see Table 6.18), and contradicts the intraday MCAAR results (see Table 6.17). Still more investigation would be needed to generalize this finding, however, these results provide the ability to test and compare different sentiment filtration techniques, across different financial markets.

6.4.4.4 Impact results visualizations

In this case study, the analyst was able to visualise the impact studies results' text files using an R software script. Figure 6.3 and Figure 6.4 show two samples of the impact results for study number 3. The figures show strong reaction around the news time (window 0 to 5 minute) in intraday MCAAR, LBM and price jump statistics.

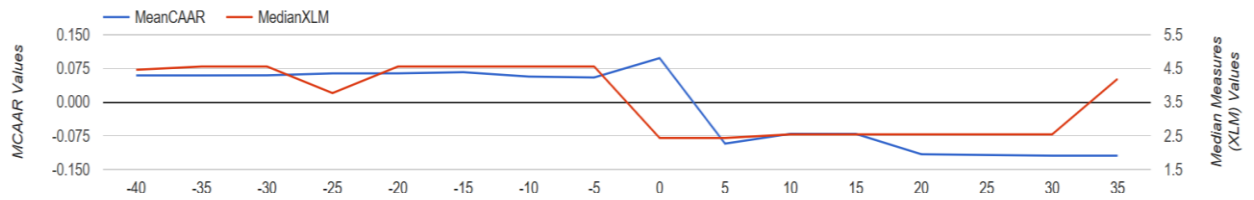


Figure 6.3 Intraday MCAAR and LBM Results for study No. 3

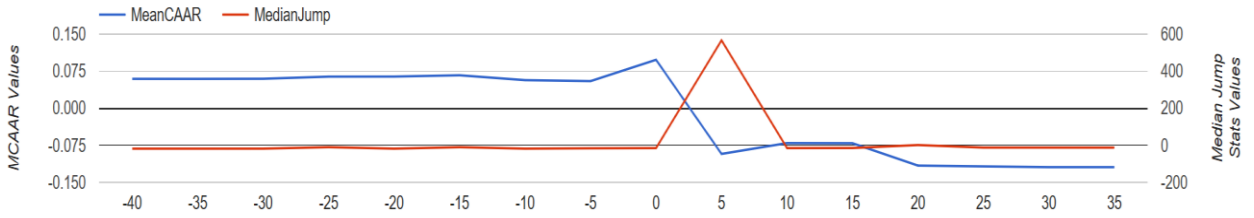


Figure 6.4 Intraday MCAAR and price jumps statistics Results for study No. 3

6.5 Discussion

This chapter has described three case studies of using the proposed NSIA framework for conducting realistic impact studies using real market data and a commercial sentiment dataset. The results of the case studies demonstrate that the framework has provided many interesting insights for the analyst:

- The first case study's results were simple, as only eight impact studies were conducted on two companies listed on the Australian Stock Exchange (see subsection 6.2.3). The results show that there was a weak reaction to news sentiment filtration technique (ESE_T1 algorithm), which naively considers negative news for a day as extreme if they pass a certain threshold, ignoring the possible effect same day related positive news stories could have on the results. The results improved slightly when implementing a better extreme sentiment extraction algorithm, which required the difference between the negative and positive news stories to pass a certain threshold (ESE_T2 algorithm).
- The second case study consisted of a larger number of impact studies (twenty-four studies). Companies from eight different financial market indices in four different countries were selected. The results (see subsection 6.3.3) provided stronger evidence that the choice of financial context does impact the results. The choice of sentiment extraction algorithm also varied, and the ESE_VOL algorithm provided better results than the ESE_TOT algorithm.

- The third case study focused on improving impact analysis by allowing three intraday impact models (on 5-minute interval timeseries) to be used (see subsection 6.4.4). As expected, the impact results were much more significant than those of the previous case studies. They show the immediate impact of releasing the news with a short time period (30 minutes after releasing the negative news), where significant drops in returns or rise of trading costs were observed. Except for few cases (e.g. impact studies related to Germany), the three impact models provided similar results. The intraday price jump statistics, complemented the intraday mean cumulative average abnormal returns results, showing immediate significant abnormal price statistics variations after releasing the news as opposed to before releasing the news, however, this impact model wasn't as sensitive to news as the other two.

In addition, the results enabled the comparison between daily and intraday impact magnitude. Table 6.20 compares the *daily* MCAAR and *intraday* MCAAR impact figures. Column Intraday MCAAR Impact Magnitude show the difference between MCCAR (0 min, 30 min) figures and MCAAR (-40 min, -10 min) figures, generated in the third case study. Column P-value represents the statistical significance of the figures shown in the Intraday MCAAR Impact Magnitude column, Column Daily MCAAR Impact Magnitude show the daily figures generated in the second case study. Like the P-value column, column Generalized Sign Z determines the statistical significance of the daily impact figures. The statistical significance is highlighted with symbols $, *, **, and ***, which denote the statistical significance at the 0.10, 0.05, 0.01 and 0.001 levels, respectively.

Table 6.20 Intraday vs Daily MCAAR results

Study no	Country	Benchmark	ESE Algo.	Intraday Impact Results					Daily Impact Results	
				MCAAR (-40 min, -10 min)	MCAAR (0 min, 30 min)	Intraday MCAAR Impact Magnitude	Welch Two Sample T-Test	P Value	Daily MCAAR Impact Magnitude	Generalized Sign Z
1	Australia	ATLI	ESE_VOL	3.65%	4.47%	0.82%	-1.49	0.9067	-0.03%	-0.216
2	Australia	ATLI	ESE_TOT	1.12%	2.18%	1.06%	-3.74	0.9979	-1.05%	-1.870*

3	Australia	AORD	ESE_VOL	6.13%	-6.97%	-13.10%	4.51	0.00201**	-0.25%	0.265
4	Australia	AORD	ESE_TOT	1.23%	0.38%	-0.85%	6.79	0.00006***	-0.82%	-1.383$
5	Germany	GDAXI	ESE_VOL	-25.23%	-24.79%	0.44%	-4.14	0.9992	-0.31%	0.926
6	Germany	GDAXI	ESE_TOT	3.66%	3.7%	0.04%	-0.83	0.7857	0.28%	0.622
7	Canada	SPTSE	ESE_VOL	2.99%	2.47%	-0.52%	3.04	0.01431**	-1.23%	-0.874
8	Canada	SPTSE	ESE_TOT	2.33%	1.54%	-0.79%	8.89	0.00002***	-0.07%	1.353$
9	USA	DJI	ESE_VOL	4.53%	3.29%	-1.24%	6.39	0.00008***	-0.10%	0.575
10	USA	DJI	ESE_TOT	9.63%	9.45%	-0.18%	2.13	0.06031$	0.69%	-0.265
11	USA	HWI	ESE_VOL	7.21%	5.86%	-1.35%	4.99	0.00101***	0.32%	2.953**
12	USA	HWI	ESE_TOT	8.87%	8.63%	-0.24%	1.77	0.07433$	0.95%	1.482$

In the light of the results, several observations can be summarised as follows:

- The table confirms that daily MCAAR figures show weaker signs of reaction to the negative news sets identified as extreme.

- There are cases in which daily figures don't reveal any impact at all. For example, in studies 9 and 10 related to Dow Jones Index, the impact of the daily MCAAR figures showed no signs of reaction to news, while the intraday impact magnitude and the p-values showed more reliable immediate impact of news on the abnormal returns. The daily MCAAR impact magnitude and Generalized Sign Z columns show the markets absorbed the negative news and by the end of the day the abnormal returns evaporated as compared to time periods around the news time.
- There are cases where the intraday MCAAR figures showed no signs of reaction to negative sentiment in the news, while the daily impact figures showed negative MCAAR figures e.g. studies 2 and 5 related to Australia's ASX top 20 leaders and Germany's DAX Index respectively. We can't attribute the impact in these cases to the negative news, as the intraday figures showed no signs of impact, this could mean the daily figures are just by chance, or there could be other factors contributing to these results. In conclusion, relying on the daily impact results in these cases could be misleading and not a true reflection of the real impact.

6.6 Conclusion

This chapter conducted three case studies, which enabled evaluating the NSIA framework using various financial contexts, sentiment extraction and impact measures parameters. The next chapter will conclude this thesis and outline the main contributions, findings, limitations of this thesis.

7 CONCLUSION AND FUTURE WORK

The key contribution of this thesis lies in proposing a framework called News Sentiment Impact Analysis (NSIA). The framework enables evaluating the impact of sentiment analysis datasets on financial markets in a flexible, consistent, and systematic manner. In this chapter, section 7.1 summarises the thesis work and findings, then section 7.2 discusses how the research questions have been addressed. Next, the research benefits are presented in section 7.3 along with the thesis limitations in section 7.4. Finally, future work is outlined in section 7.5.

7.1 Thesis summary

This thesis has investigated the area of sentiment analysis and its impact on financial market entities. The thesis started with reviewing sentiment analysis processes and techniques, describing the various steps taken to convert text corpuses into sentiment metrics. The review mentioned with brief detail some of the commercially available sentiment datasets, which mainly analyse news articles related to financial markets around the world. Next, the review described the various concepts related to market data as well as the various market measures affected by news sentiment. This was followed by an overview of existing studies related to sentiment analysis and finance, illustrating the sentiment data, financial context, market data, and impact models that have been used in these studies. The literature shows a gap in defining systematic and reusable evaluation processes, that could be used by a wide range of users to automatically conduct impact analysis of sentiment datasets in different financial contexts.

In chapter 3, the thesis raises the following three research questions:

- What are the parameters that uniquely define a context that enables validating sentiment datasets to conduct impact studies in multiple financial contexts?
- Given a context, what is the set of use cases that needs to be defined to guide users to conduct their experiments in a consistent fashion?

- How to support automating the use cases identified in the second research question within a software framework?

In answering these research questions, the thesis proposes a framework called "News Sentiment Impact Analysis (NSIA)". It consists of three components which are a novel conceptual data model, a software architecture, and a set of use cases. In chapter 4, the key component which is the conceptual data model (also called Comparison Parameters Data (CPD) model) is described. It captures three sets of parameters that enable conducting a wide range of sentiment-driven impact analysis studies. The first set consists of the financial context parameters, which enable the analyst (any user with knowledge in financial markets) to select any financial markets context, in which financial entities and benchmarks are defined. The second set consists of sentiment related parameters, which capture the attributes found within a sentiment dataset, and define a number of sentiment filtration and extreme news extraction algorithms. The third set consists of the impact measure parameters, which enable the analyst to choose the impact model and set the threshold variables related to each impact model. The CPD model is supported by a software architecture, which consists of a GUI, Business and Data layers. The main use case allows the analyst to define the CPD parameters using the provided GUI layer and trigger the impact analysis process.

To evaluate the effectiveness of the proposed framework, a prototype described in chapter 5 was implemented. The prototype integrates a number of software tools and packages that fit the requirements depicted by the NSIA framework. Chapter 6 discusses three case studies that have been conducted to evaluate the different aspects of the proposed framework, using real market data and a commercially available sentiment dataset.

7.2 Addressing the research questions

This section discusses how the proposed framework (NSIA framework) has addressed the research questions.

- The first research question "What are the parameters that uniquely define a context that enables validating sentiment datasets by conducting impact studies in multiple financial contexts?" has been addressed by proposing a novel data model called the Comparison Parameters Data (CPD) model. The CPD model allows multiple financial contexts to be defined, various sentiment filtration and extreme extraction methods to be selected, and a number of impact measures to be

applied. The CPD model has been evaluated using three case studies, which demonstrated the flexibility of the proposed data model to define multiple financial contexts, sentiment, and impact measure parameters.

- The second research question "Given a context, what are the set of use cases that need to be defined to guide users to conduct their experiments in a consistent fashion?" has been addressed by proposing three use cases, which guide the user to import market and sentiment data into the CPD model and conduct impact analysis studies. The three case studies demonstrated that it was possible to conduct sixty distinct impact studies and consistently evaluate the impact for each study using repeatable and consistent steps.
- The third research question *"How to support automating the use cases identified in the second research question within a single software framework?"* has been addressed via the software architecture. The case studies demonstrated that a prototype implementation of the architecture has been able to integrate different software packages (e.g. Eventus) and libraries (e,g, R) to handle operations from importing market data from different sources, to filtering news sentiment datasets in the database, to conducting statistical testing, and delivering impact results and visualizations.

7.3 Benefits of the research

The research conducted in this thesis has many benefits such as:

- The automation of impact analysis has other applications besides conducting evaluations of sentiment datasets. For example, it can be used to validate algorithmic trading strategies. The analyst can conduct a series of impact analysis studies to fine tune the parameters of a particular algorithmic trading strategy or to evaluate the performance of a strategy in a particular financial context.
- Impact analysis could also be performed on a large scale for example, scripts could be put in place to automatically define thousands of financial contexts and conduct impact analysis in these contexts. This automation could for instance provide the analyst with insights on what is the most suitable context for a particular sentiment dataset.

- From another angle, the idea behind the framework is applicable for automating impact analysis in other domains. For instance, evaluating the impact of different sentiment measures of social networks data based on different contextual parameters (e.g. parameters that define marketing contexts instead of financial contexts). Other domains include for instance, evaluating the impact of sentiment datasets on peoples' decision making for example in buying/selling products and services (Pang et al., 2002; Jebaseeli & Kirubakaran, 2012; Dhaoui et al., 2017), in planning their holiday destinations (Cruz, et al.,2013), or digital advertising (Yang et al., 2016).

7.4 Thesis limitations

This section discusses the main limitations of the proposed framework and this thesis. The case studies results revealed many limitations of the Comparison Parameters Data (CPD) model in terms of modelling any context for example:

- A parameter to define the country (where stock exchanges are based), as it became apparent from conducting the studies, that financial impact does in fact vary a lot between countries.in particular if there are differences between the language(s) used in the sentiment dataset and language(s) used in the county.
- A parameter to define the company size, as impact varies according to the size of the company (large, medium or small capital). For example, an impact model considering liquidity as the impact measure shows liquidity varied between companies according to their capital size (Gomber et al., 2015).
- There is a limited number of impact measures which use regression analysis to evaluate impact. The CPD model doesn't support yet trading strategies as a technique to evaluate impact.

In addition, the evaluation processes are not fully automated. There are no use cases to support automated importing of market and sentiment datasets into the CPD model. Especially if large scale impact analysis studies are going to be conducted, new use cases should be in place to support importing data on the fly. The existing use cases have utilized techniques that require technical knowledge such as Unix and PL/SQL scripts to import market and sentiment data. Moreover, Extreme Sentiment Extraction (ESE) algorithms are all prefixed. Defining new ESE algorithms still needs an IT expert to implement them. Users can't define their own, as it requires programming skills.

Finally, the NSIA framework has only validated in one sentiment dataset which has been imported from Thomson Reuters. This is due to time limitations which prevented the use of other sentiment datasets.

7.5 Future work

This section describes some potential research avenues and future work, which have been flagged as essential steps towards more mature methodologies for conducting sentiment driven impact analysis. In the short term, future work can focus on the following:

- Extending the CPD model to introduce additional parameters that give a more precise representation of a financial context such as country and company size. Extending the model would provide a better understanding of the impact results as it would relate them to a more accurate context.
- Extending the architecture with capabilities to enable the analyst (without technical knowledge) to define some of the complex parameters such as filtration and extreme sentiment extraction algorithms. Some visual notation or a representation based on mathematics could be a solution.
- Introducing new use cases to increase automation and allow large scale impact studies to be conducted. Such use cases can provide end to end analytics which would include importing market and sentiment data on the fly, evaluating thousands of financial contexts against one sentiment dataset, and allow the user to interact visually with the results.
- Incorporating additional sentiment datasets into the framework. Like for example the sentiment datasets offered by Quandl (2016) and Ravenpack (2016). It would be interesting to provide the users with the ability to compare impact results of more than one sentiment data source, which could bring better understanding how different financial contexts respond to these datasets.

On the longer run, the framework can be further collectively enhanced by a community of researchers, and made publicly available to researchers, as a generic framework for automating realistic and complex impact analysis studies.

8 REFERENCES

Agrawal, M., Kishore, R., & Rao, H. R. (2006). Market reactions to e-business outsourcing announcements: An event study. *Information & Management, 43*(7), 861-873.

Agence France-Presse. (2016). *Agence france-presse.* Retrieved April, 2016, from http://www.afp.com/en/home/

AlchemyAPI. (2016). Retrieved May, 2016, from http://blog.mashape.com/list-of-20-sentiment-analysis-apis/

Allen, D. E., McAleer, M., & Singh, A. K. (2015). *Daily Market News Sentiment and Stock Prices* (No. 15-090/III). Tinbergen Institute Discussion Paper.

Anderson, D. L. (2000). Management information systems: Solving business problems with information technology McGraw-Hill, Inc.

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, *59*(3), 1259-1294.

Apache UIMA. (2016). *Apache UIMA.* Retrieved April, 2016, from http://uima.apache.org

Associated press. (2016). *Associated press.* Retrieved April, 2016, from http://www.ap.org/

Atwell, E. *The brown corpus tag-set.* Retrieved April, 2016, from https://www.comp.leeds.ac.uk/ccalas/tagsets/brown.html

Australian Stock Exchange (ASX). (2014). *All ordinaries index.* Retrieved September, 2014, from http://www.asx.com.au/listings/listing-IPO-on-ASX.htm

Azar, P. D. (2009). *Sentiment analysis in financial news* (Doctoral dissertation, Harvard University).

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, *61*(4), 1645-1680.

Bloomberg. (2016). *Bloomberg news and stocks data feed.* Retrieved April, 2016, from http://www.bloomberg.com/markets/stocks

Bloomberg Sentiment Data. (2016). *Sentiment analysis of financial news and social media.* Retrieved April 2016, from http://www.bloomberglabs.com/data-science/projects/sentiment-analysis-financial-news-social-media/

Bohn, N., Rabhi, F. A., Kundisch, D., Yao, L., & Mutter, T. (2012). Towards automated event studies using high frequency news and trading data. In *International Workshop on Enterprise Applications and Services in the Finance Industry* (pp. 20-41). Springer, Berlin, Heidelberg.

Bollen, J., & Mao, H. (2011). Twitter mood as a stock market predictor. *Computer*, *44*(10), 91-94.

Brants, T. (2000). TnT: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing* (pp. 224-231). Association for Computational Linguistics.

Brown, S. J., & Warner, J. B. (1985). Using daily stock returns: The case of event studies. *Journal of financial economics*, *14*(1), 3-31.

Business wire. (2016). *Business wire about.* Retrieved April, 2016, from http://www.businesswire.com/portal/site/home/about/

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, *28*(2), 15-21.

Cambria, E., Song, Y., Wang, H., & Howard, N. (2014). Semantic multidimensional scaling for open-domain sentiment analysis. *IEEE Intelligent Systems*, *29*(2), 44-51.

Cambria, E., Xia, Y., & Hussain, A. (2012). Affective Common-Sense Knowledge Acquisition for Sentiment Analysis. In *LREC* (pp. 3580-3585).

Chatzakou, D., Passalis, N., & Vakali, A. (2015). Multispot: Spotting sentiments with semantic aware multilevel cascaded analysis. *International Conference on Big Data Analytics and Knowledge Discovery*, pp. 337-350.

Chen, H., De, P., Hu, Y. J., & Hwang, B. H. (2013, June). Customers as advisors: The role of social media in financial markets. In *3rd Annual Behavioural Finance Conference, Queen's University, Kingston, Canada.< http://www. bhwang. com/customers. pdf*.

Choi, J. W. (2013). *Reason #2 why markets are not efficient: People are emotional.* Retrieved July, 2016, from https://www.moneygeek.ca/weblog/2013/06/20/debunking-markets-are-efficient-myth-part-4/

Corrado, C. J. (2011). Event studies: A methodology review. *Accounting & Finance*, *51*(1), 207-234.

Cowan Research LC, U. (2016). *Eventus software.* Retrieved February/2015 http://www.eventstudy.com/index.html

Cruz, F. L., Troyano, J. A., Enríquez, F., Ortega, F. J., & Vallejo, C. G. (2013). 'Long autonomy or long delay?'The importance of domain in opinion mining. Expert Systems with Applications, 40(8), 3174-3184.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management science*, *53*(9), 1375-1388.

Davis, A. K., Piger, J. M., & Sedor, L. M. (2012). Beyond the numbers: Measuring the information content of earnings press release language. *Contemporary Accounting Research*, *29*(3), 845-868.

Davis, A. K., & Tama-Sweet, I. (2012). Managers' use of language across alternative disclosure outlets: Earnings press releases versus MD&A. *Contemporary Accounting Research*, *29*(3), 804-837.

Davis, A. K., Ge, W., Matsumoto, D., & Zhang, J. L. (2015). The effect of manager-specific optimism on the tone of earnings conference calls. *Review of Accounting Studies*, *20*(2), 639-673.

Demers, E. A., & Vega, C. (2014). Understanding the role of managerial optimism and uncertainty in the price formation process: Evidence from the textual content of earnings announcements.

Digitext. (2017). *DICTION: Text analysis tool.* Retrieved August, 2017, from http://www.dictionsoftware.com/

Dhaoui, C., Webster, C., & Tan, L. P. (2017). Social media sentiment analysis: lexicon versus machine learning. *Journal of Consumer Marketing*, (just-accepted), 00-00.

Doran, J. S., Peterson, D. R., & Price, S. M. (2012). Earnings conference call content and stock price: the case of REITs. *The Journal of Real Estate Finance and Economics*, *45*(2), 402-434.

Dzielinski, M. (2011). News sensitivity and the cross-section of stock returns. *Available at SSRN*.

EDGAR. (2017). *EDGAR online.* Retrieved July, 2017, from http://www.edgar-online.com/

Engle, R. (2001). GARCH 101: The use of ARCH/GARCH models in applied econometrics. *The Journal of Economic Perspectives*, *15*(4), 157-168.

Engelberg, J. (2008). Costly information processing: Evidence from earnings announcements.

Engelberg, J. E., Reed, A. V., & Ringgenberg, M. C. (2012). How are shorts informed?: Short sellers, news, and information processing. *Journal of Financial Economics*, *105*(2), 260-278.

Esuli, A., & Sebastiani, F. (2006). SENTIWORDNET: A high-coverage lexical resource for opinion mining. *Institute of Information Science and Technologies (ISTI) of the Italian National Research Council (CNR).*

Feldman, R., Govindaraj, S., Livnat, J., & Segal, B. (2008). The incremental information content of tone change in management discussion and analysis.

Ferguson, N. J., Philip, D., Lam, H. Y., & Guo, J. M. (2015). Media content and stock returns: The predictive power of press.

Feuerriegel, S., & Neumann, D. (2014). Evaluation of news-based trading strategies. *International Workshop on Enterprise Applications and Services in the Finance Industry,* pp. 13-28.

French, K. R. (1980). Stock returns and the weekend effect. *Journal of financial economics, 8*(1), 55-69.

Frost, J. (2016). *Understanding t-tests: T-values and t-distributions.* Retrieved June, 2017, from http://blog.minitab.com/blog/adventures-in-statistics-2/understanding-t-tests-t-values-and-t-distributions

GATE. (2016). *Gate software.*, Retrieved June, 2016, from https://gate.ac.uk/

General Inquirer. (2016). *General inquirer.*, 2016, from http://www.wjh.harvard.edu/~inquirer/

Gershenson, C. (2003). Artificial neural networks for beginners. *arXiv preprint cs/0308031*.

Glance, N., Hurst, M., Nigam, K., Siegler, M., Stockton, R., & Tomokiyo, T. (2005). Deriving marketing intelligence from online discussion. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 419-428). ACM.

Gomber, P., Schweickert, U., & Theissen, E. (2015). Liquidity dynamics in an electronic open limit order book: An event study approach. *European Financial Management, 21*(1), 52-78.

Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Use of sentiment analysis for capturing patient experience from free-text comments posted online. *Journal of medical Internet research, 15*(11).

Gupta, V., & Lehal, G. S. (2013). A survey of common stemming techniques and existing stemmers for indian languages. *Journal of Emerging Technologies in Web Intelligence, 5*(2), 157-161.

Harcar, D. M. *Justification and expected benefits of data analysis automation projects.* Retrieved August, 2016, from https://www.statsoft.com/Portals/0/Support/Download/White-Papers/Automation-Projects.pdf

Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems, 55*(3), 685-697.

Hesham, A. (2017). *Advantages and disadvantages of DBMS over traditional file processing system?* Retrieved May, 2017, from https://www.bayt.com/en/specialties/q/47871/advantages-and-disadvantages-of-dbms-over-traditional-file-processing-system/

Hong, Y., & Skiena, S. (2010). The Wisdom of Bookies? Sentiment Analysis Versus. the NFL Point Spread. In *ICWSM*.

Henry, E. (2008). Are investors influenced by how earnings press releases are written?. *The Journal of Business Communication (1973)*, *45*(4), 363-407.

Henry, E., & Leone, A. J. (2009). Measuring qualitative information in capital markets research.

Huang, C. J., Liao, J. J., Yang, D. X., Chang, T. Y., & Luo, Y. C. (2010). Realization of a news dissemination agent based on weighted association rules and text mining techniques. *Expert Systems with Applications*, *37*(9), 6409-6413.

Huang, X., Teoh, S. H., & Zhang, Y. (2013). Tone management. *The Accounting Review*, *89*(3), 1083-1113.

Inkpen, D., & Désilets, A. (2005, October). Semantic similarity for detecting recognition errors in automatic speech transcripts. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 49-56). Association for Computational Linguistics.

Investopedia. (2016a). *Macroeconomic announcements.* Retrieved April, 2016, from http://www.investopedia.com/terms/m/macroeconomics.asp

Investopedia. (2016b). *Liquidity.* Retrieved April, 2016, from http://www.investopedia.com/terms/l/liquidity.asp?o=40186&l=dir&qsrc=999&qo=investopediaSiteSearch&ad=SEO&ap=google.com.au&an=SEO

InvestingAnswers. (2017a). *Stock quote.* Retrieved June, 2017, from http://www.investinganswers.com/financial-dictionary/stock-market/stock-quote-5154

InvestingAnswers. (2017b). *Market index.* Retrieved June, 2017, from http://www.investinganswers.com/financial-dictionary/investing/market-index-1305

Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. (2011). Data replication & reproducibility. again, and again, and again .... introduction. *Science (New York, N.Y.)*, *334*(6060), 1225.

Jakob, N., & Gurevych, I. (2010). Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 1035-1045). Association for Computational Linguistics.

Jebaseeli, A. N., & Kirubakaran, E. (2012). A survey on sentiment analysis of (product) reviews. *International Journal of Computer Applications*, *47*(11).

Jegadeesh, N., & Wu, D. (2013). Word power: A new approach for content analysis. *Journal of Financial Economics*, *110*(3), 712-729.

Just Data. (2017). *Track comprehensive end-of-day data for ASX and other stock exchanges, with BodhiGold.* Retrieved June, 2017, from http://www.justdata.com.au/asx-end-of-day-eod-data.php

Kothari, S., & Warner, J. B. (2004). The econometrics of event studies.

Kothari, S. P., Li, X., & Short, J. E. (2009). The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. *The Accounting Review*, *84*(5), 1639-1670.

Kouloumpis, E., Wilson, T., & Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg!. *Icwsm*, *11*(538-541), 164.

Lee, S. S., & Mykland, P. A. (2007). Jumps in financial markets: A new nonparametric test and jump dynamics. *The Review of Financial Studies*, *21*(6), 2535-2563.

Lexalytics. (2016). Retrieved May, 2016, from https://www.lexalytics.com/

Li, F. (2010). The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach. *Journal of Accounting Research*, *48*(5), 1049-1102.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

Lo, A. W. (2004). The adaptive markets hypothesis. *The Journal of Portfolio Management*, *30*(5), 15-29.

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

Lugmayr, A. (2013). Predicting the future of investor sentiment with social media in stock exchange investments: A basic framework for the DAX performance index. In *Handbook of social media management* (pp. 565-589). Springer Berlin Heidelberg.

Lugmayr, A., & Gossen, G. (2013). Evaluation of Methods and Techniques for Language Based Sentiment Analysis for DAX 30 Stock Exchange A First Concept of a "LUGO" Sentiment Indicator. *International SERIES on Information Systems and Management in Creative eMedia*, (1), 69-76.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* Cambridge university press Cambridge.

Mallet. (2016). *MAchine learning for LanguagE toolkit.*, Retrieved July, 2016, from http://mallet.cs.umass.edu/

Martin, R. (2014). *Understanding sentiment analysis and sentiment accuracy.* Retrieved April, 2016, from http://blog.infegy.com/understanding-sentiment-analysis-and-sentiment-accuracy

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093-1113.

Milosevic, Z., Chen, W., Berry, A., & Rabhi, F. A. (2016). An open architecture for event-based analytics. *International Journal of Data Science and Analytics*, *2*(1-2), 13-27.

Milton, A. (2016). *Order book, level 2 market data and depth of market.* Retrieved June, 2017, from https://www.thebalance.com/order-book-level-2-market-data-and-depth-of-market-1031118

Mitra, G., & Mitra, L. (Eds.). (2011). *The handbook of news analytics in finance* (Vol. 596). John Wiley & Sons.

Mittermayer, M. A. (2004, January). Forecasting intraday stock price trends with text mining techniques. In *system sciences, 2004. proceedings of the 37th annual hawaii international conference on* (pp. 10-pp). IEEE.

Murch, R. (2001). Project management: Best practices for IT professionals. Prentice Hall Professional.

Nordquist, R. (2016). *Corpus (language).* Retrieved May, 2017, from https://www.thoughtco.com/what-is-corpus-language-1689806

Nicholls, C., & Song, F. (2010). Comparison of feature selection methods for sentiment analysis. *Advances in Artificial Intelligence*, 286-289.

Niederhoffer, V. (1971). The analysis of world events and stock prices. *The Journal of Business*, *44*(2), 193-219.

NLTK. (2016). Learning to classify text. Retrieved Feb, 2016, from http://www.nltk.org/book/ch06.html

Open NLP. (2016). Apache OpenNLP., Retrieved Feb, 2016, from http://opennlp.apache.org

Oracle Corporation. (2017a). *Oracle SQL developer tool.* Retrieved May, 2017, from http://www.oracle.com/technetwork/developer-tools/sql-developer/downloads/index.html

Oracle Corporation. (2017b). *Oracle PL/SQL scripting language.* Retrieved May, 2017, from http://www.oracle.com/technetwork/database/features/plsql/index.html

Oracle Corporation. (2017c). *SQL loader.* Retrieved June, 2017, from https://docs.oracle.com/cd/B19306_01/server.102/b14215/ldr_concepts.htm

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (pp. 79-86). Association for Computational Linguistics.

Peng, R. D. (2011). Reproducible research in computational science. *Science, 334*(6060), 1226-1227.

PR newswire. (2016). *PR newswire about.* Retrieved April, 2016, from http://prnewswire.mediaroom.com/about-pr-newswire

Price, S. M., Doran, J. S., Peterson, D. R., & Bliss, B. A. (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance, 36*(4), 992-1011.

ProgrammableWeb. (2016). *Information exchange mediums.* Retrieved April, 2016, from http://www.programmableweb.com/category/News%20Services/apis?category=20250

Quandl. (2016). *Quandl AAII investor sentiment data.* Retrieved April, 2016, from https://www.quandl.com/data/AAII/AAII_SENTIMENT-AAII-Investor-Sentiment-Data

Rahman, A. (2014). *Behavioural models & sentiment analysis applied to finance conference 2014.* Retrieved April, 2016, from http://www.optirisk-systems.com/blog/index.php/behavioural-models-sentiment-analysis-applied-to-finance-conference-2014/

R Software. (2017). *The R project for statistical computing.* Retrieved May, 2017, from https://www.r-project.org/

Rabhi, F. A., Guabtni, A., & Yao, L. (2009). A data model for processing financial market and news data. *International Journal of Electronic Finance, 3*(4), 387-403.

Rabhi, F. A., Yao, L., & Guabtni, A. (2012). ADAGE: A framework for supporting user-driven ad-hoc data analysis processes. *Computing, 94*(6), 489-519.

RavenPack. (2016). RavenPack. Retrieved April, 2016, from http://www.ravenpack.com/

RavenPack News Analytics. (2016). RavenPack news analytics: Turning text from traditional & social media into a structured data feed for quantitative applications . Retrieved April, 2016, from http://www.ravenpack.com/products/ravenpack-news-analytics/

RESTful. (2016). *RESTful web services tutorial.* Retrieved April, 2016, from http://www.tutorialspoint.com/restful/

Rice, J. (2006). Mathematical statistics and data analysis. Nelson Education.

Riordan, R., Storkenmaier, A., Wagener, M., & Zhang, S. S. (2013). Public information arrival: Price discovery and liquidity in electronic limit order markets. *Journal of Banking & Finance, 37*(4), 1148-1159.

Robertson, C. S., Rabhi, F. A., & Peat, M. (2013). A service-oriented approach towards real time financial news analysis. *Consumer Information Systems and Relationship Management: Design, Implementation, and Use: Design, Implementation, and Use*, 32.

Rockefeller, B. (2016). *How to interpret trade volume.* Retrieved March, 2016, from http://www.dummies.com/how-to/content/how-to-interpret-trade-volume.html

Runeson, P., & Höst, M. (2009). Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering*, *14*(2), 131.

Scala NLP. (2016). *Scientific computing, machine learning, and natural language processing.*, Retrieved April, 2016, from http://www.scalanlp.org/

Schumaker, R. P., Zhang, Y., Huang, C. N., & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems*, *53*(3), 458-464.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47.

SenticNet. (2014). *Semantic based sentiment analysis.* Retrieved April, 2014, from http://sentic.net/api/en/concept/celebrate_special_occasion/

Shuttleworth, M. (2016). *Case study research design.*, Retrieved April 2016, from https://explorable.com/case-study-research-design

Siering, M. (2012a). " Boom" or" Ruin"--Does It Make a Difference? Using Text Mining and Sentiment Analysis to Support Intraday Investment Decisions. In *System Science (HICSS), 2012 45th Hawaii International Conference on* (pp. 1050-1059). IEEE.

Siering, M. (2012b). Investigating the impact of media sentiment and investor attention on financial markets. In *International Workshop on Enterprise Applications and Services in the Finance Industry* (pp. 3-19). Springer, Berlin, Heidelberg.

Sirca. (2017). *Thomson Reuters Tick History portal.* Retrieved June, 2017, from https://tickhistory.thomsonreuters.com/TickHistory/login.jsp

Stanford NLP. (2016). *Stanford natural language processing*, Retrieved June, 2016, from http://nlp.stanford.edu/

Statistics Solutions. (2017). *How to conduct the wilcoxon sign test.* Retrieved June, 2017, from http://www.statisticssolutions.com/how-to-conduct-the-wilcox-sign-test/

Stonebank, M. (2000). *UNIX introduction.* Retrieved May, 2017, from http://www.ee.surrey.ac.uk/Teaching/Unix/unixintro.html

StreamHacker. (2010). *Text classification for sentiment analysis – precision and recall.* Retrieved April, 2016, from http://streamhacker.com/2010/05/17/text-classification-sentiment-analysis-precision-recall/

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*(3), 1139-1168.

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, *63*(3), 1437-1467.

Thomson Reuters. (2014). *Thomson Reuters News Analytics(TRNA).* Retrieved Jan, 2014, from http://thomsonreuters.com/products/financial-risk/01_255/news-analytics-product-brochure--oct-2010.pdf

Tsay, R. S. (2005). *Analysis of financial time series* (Vol. 543). John Wiley & Sons.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *Icwsm*, *10*(1), 178-185.

Twitter About. (2016). Retrieved March, 2016, from https://about.twitter.com/company

Viney, C. (2003). Financial institutions, instruments and markets. McGraw-Hill.

Vu, T. T., Chang, S., Ha, Q. T., & Collier, N. (2012). An experiment in integrating sentiment features for tech stock prediction in twitter.

Wu, Y., Zhang, Q., Huang, X., & Wu, L. (2009). Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3* (pp. 1533-1541). Association for Computational Linguistics.

Yang, S., Lin, S., Carlson, J. R., & Ross Jr, W. T. (2016). Brand engagement on social media: will firms' social media efforts influence search engine advertising effectiveness?. *Journal of Marketing Management*, *32*(5-6), 526-557.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Icml* (Vol. 97, pp. 412-420).

Yao, L., & Rabhi, F. A. (2015). Building architectures for data-intensive science using the ADAGE framework. *Concurrency and Computation: Practice and Experience*, *27*(5), 1188-1206.

Zamora, A. (2016). *Modifications to the lancaster stemming algorithm.* Retrieved April, 2016, from http://www.scientificpsychic.com/paice/paice.html

ZetCode. (2017). *Java swing tutorial.* Retrieved February, 2016, from http://zetcode.com/tutorials/javaswingtutorial/

Zhang, L. (2013). Sentiment analysis on Twitter with stock price and significant keyword correlation (Doctoral dissertation).

Zhou, X., Tao, X., Yong, J., & Yang, Z. (2013, June). Sentiment analysis on tweets for social events. In *Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on* (pp. 557-562). IEEE.

APPENDIX A: ADDITIONAL INFORMATION FOR NSIA FRAMEWORK

Impact study parameters are preserved into IMPACT_STUDIES_LOG database object. Sample impact study parameters are shown in Figure A. 1.

STUDY...	RUN_DATE	PARAMETER_NAME	PARAMETER_VALUE	PARAMETER_TYPE
1	23/MAR/16	Entity E	ANZ.AX	Fin. Context
1	23/MAR/16	Entity E	QAN.AX	Fin. Context
1	23/MAR/16	Impact_Analysis_Tool	Eventus	Impact Measure
1	23/MAR/16	Estimation_Window	30 Days before and...	Impact Measure
1	23/MAR/16	Impact_Window	1 Day before and a...	Impact Measure
1	23/MAR/16	Impact_Measure	Daily MCAR	Impact Measure
1	23/MAR/16	ESE_ALGO	ESE_VOL	Sentiment (ESE)
1	23/MAR/16	Num_co	1	Sentiment (Fn)
1	23/MAR/16	Lnkd_cnt5	0	Sentiment (Fn)
1	23/MAR/16	Topic_code	OIL	Sentiment (Fn)
1	23/MAR/16	Item_type	Alert	Sentiment (Fn)
1	23/MAR/16	Sentiment_class	-1	Sentiment (Fn)
1	23/MAR/16	Relevance	1	Sentiment (Fn)
1	23/MAR/16	Story_Date	01/Jan/2010 to 31/...	Sentiment (Fn)
1	23/MAR/16	stock_ric	"ANZ.AX" or "QAN.A...	Sentiment (Fn)
1	23/MAR/16	Sentiment Dataset Name	TRNA	Sentiment (Fn)
1	23/MAR/16	Study Period P	01/Jan/2010 to 31/...	Fin. Context
1	23/MAR/16	Benchmark Variable BV	EOD prices	Fin. Context
1	23/MAR/16	Entity Variable EV	EOD prices	Fin. Context
1	23/MAR/16	Benchmark B	.AORD	Fin. Context
1	23/MAR/16	Entity E	BHP.AX	Fin. Context
2	23/MAR/16	Entity E	ANZ.AX	Fin. Context
2	23/MAR/16	Impact_Analysis_Tool	Eventus	Impact Measure
2	23/MAR/16	Impact_Window	1 Day before and a...	Impact Measure
2	23/MAR/16	Impact_Measure	Daily MCAR	Impact Measure
2	23/MAR/16	ESE_ALGO	ESE_TOT	Sentiment (ESE)
2	23/MAR/16	Num_co	1	Sentiment (Fn)
2	23/MAR/16	Lnkd_cnt5	0	Sentiment (Fn)
2	23/MAR/16	Topic_code	OIL	Sentiment (Fn)
2	23/MAR/16	Item_type	Alert	Sentiment (Fn)
2	23/MAR/16	Sentiment_class	-1	Sentiment (Fn)
2	23/MAR/16	Relevance	1	Sentiment (Fn)
2	23/MAR/16	Story_Date	01/Jan/2010 to 31/...	Sentiment (Fn)
2	23/MAR/16	stock_ric	"ANZ.AX" or "QAN.A...	Sentiment (Fn)
2	23/MAR/16	Sentiment Dataset Name	TRNA	Sentiment (Fn)
2	23/MAR/16	Study Period P	01/Jan/2010 to 31/...	Fin. Context
2	23/MAR/16	Benchmark Variable BV	EOD prices	Fin. Context
2	23/MAR/16	Entity Variable EV	EOD prices	Fin. Context
2	23/MAR/16	Benchmark B	.AORD	Fin. Context

Figure A. 1 Impact studies parameters and values log

The physical structure of INTRADAY_PRICE_JUMPS database object is shown in Figure A. 2. Attributes *JUMPSTATISTIC*, *ISABNORMAL* and *NEWS_ABNORMAL* store the price jump statistic relevant to the intraday time series record.

```
CREATE TABLE "INTRADAY"."INTRADAY_PRICE_JUMPS"
   (  "SECURITY" VARCHAR2(26 BYTE),
      "DATEOFTRADE" DATE,
      "STARTPERIOD" VARCHAR2(26 BYTE),
      "ENDPERIOD" VARCHAR2(26 BYTE),
      "LSTPRICE" NUMBER(12,4),
      "LSTRETURN" NUMBER(20,12),
      "TOTVOL" NUMBER(12,2),
      "TOTVAL" NUMBER(12,3),
      "NUMTRADE" NUMBER(10,2),
      "LSTASK" NUMBER(16,4),
      "LSTBID" NUMBER(16,4),
      "LSTSPREAD" NUMBER(12,4),
      "LSTMIDPOINT" NUMBER(14,4),
      "FSTPRICE" NUMBER(14,4),
      "PREVLSTPRICE" NUMBER(14,4),
      "HGHPRICE" NUMBER(14,4),
      "LOWPRICE" NUMBER(14,4),
      "AVEPRICE" NUMBER(16,6),
      "PREVAVEPRICE" NUMBER(14,4),
      "NUMTRADEBI" NUMBER(12,4),
      "NUMTRADESI" NUMBER(13,4),
      "TOTVOLBI" NUMBER(15,4),
      "TOTVOLSI" NUMBER(15,4),
      "TOTVALBI" NUMBER(17,5),
      "TOTVALSI" NUMBER(17,5),
      "DIFLSTPRICE" NUMBER(17,4),
      "DIFAVEPRICE" NUMBER(18,6),
      "VWAP" NUMBER(15,6),
      "NUMZERORETURN" NUMBER(12,4),
      "AMIHUD" NUMBER(24,12),
      "AMIVEST" NUMBER(18,8),
      "OUTPUTEVENTCOUNTER" NUMBER(13,5),
      "JUMPSTATISTIC" NUMBER(15,6),
      "ISABNORMAL" NUMBER(11,5),
      "NEWS_ABNORMAL" VARCHAR2(300 BYTE) );
```

Figure A. 2 *INTRADAY_PRICE_JUMPS* Physical database table structure

The unix shell script shown in Figure A. 3 invokes HGA tool to calculate intraday timeseries data. It expects tickistory trades and quoates datafile in csv format as an input. The script will then load the timeseries data computed by HGA tool into INTRADAY_HOMO_TS database object.

```
/*Task: load tick data into HGA to generate  timeseries (Note: This step  expects a
csv file named InputTRTH.csv )
    1- First Split Large CSV  TickData file imported from Thomson Reuters     into
small size files
    2- Then Loop through the these small TickData csv files and pass them one by one
       to HGA tool to calculate the Timeseries
    3- Next, load the generated timseries csv file into the database object
*/

/* 1- First Split Large CSV  TickData file imported from Thomson Reuters     into
small size files */
split -l 200000 InputTRTH.csv InputTRTH_

/*     2- Then Loop through the these small TickData csv files and pass them one by
one to HGA tool to calculate the Timeseries
*/
for FileName IN (InputTRTH_*) Loop

     /* Call HGA Tool, which will generate Timeseries files with _output suffix i.e.
InputTRTH_aa_output.csv */
     HGA_Main.exe
End Loop

/*    3- Next, load the generated timeseries csv file into the database object */
Process2_LoadTimeSeriesIntoDataBase.sh
```

Figure A. 3 Unix shell script to compute Intraday returns time series

APPENDIX B: GUI IMPLEMENTATION OF NSIA FRAMEWORK

The GUI shown in Figure B. 1 and Figure B. 2 illustrate the steps to initiate and define Financial Context parameters and save the study, which will generate a unique study number.

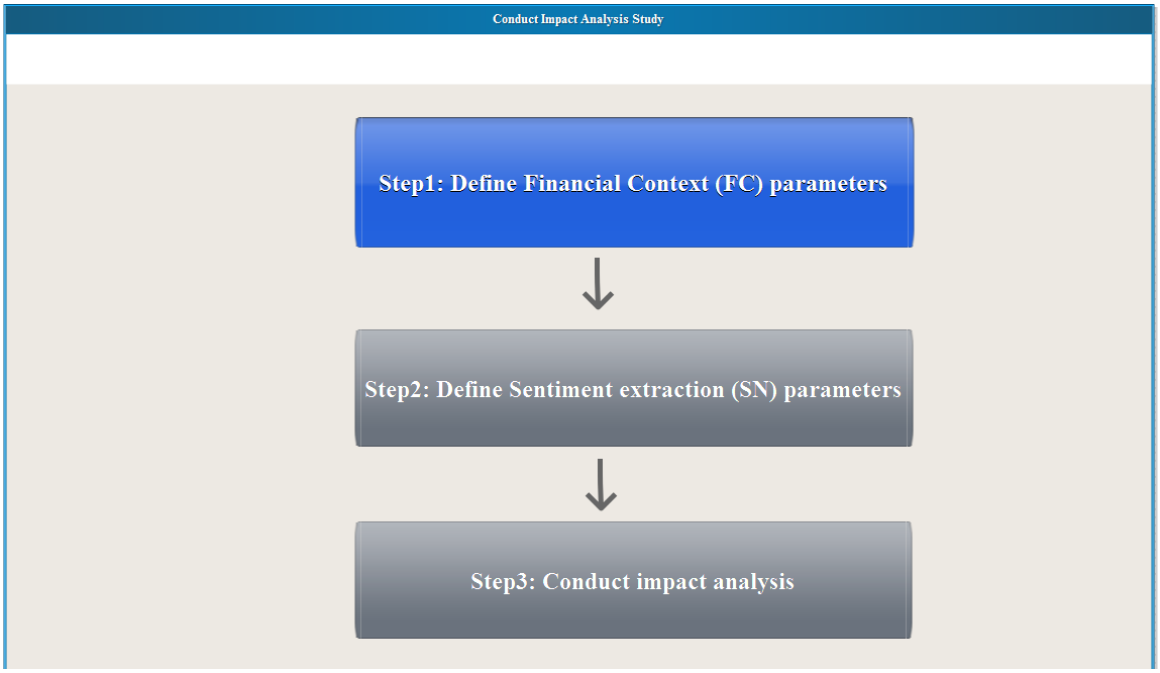


Figure B. 1 Main GUI showing steps to conduct impact t analysis study

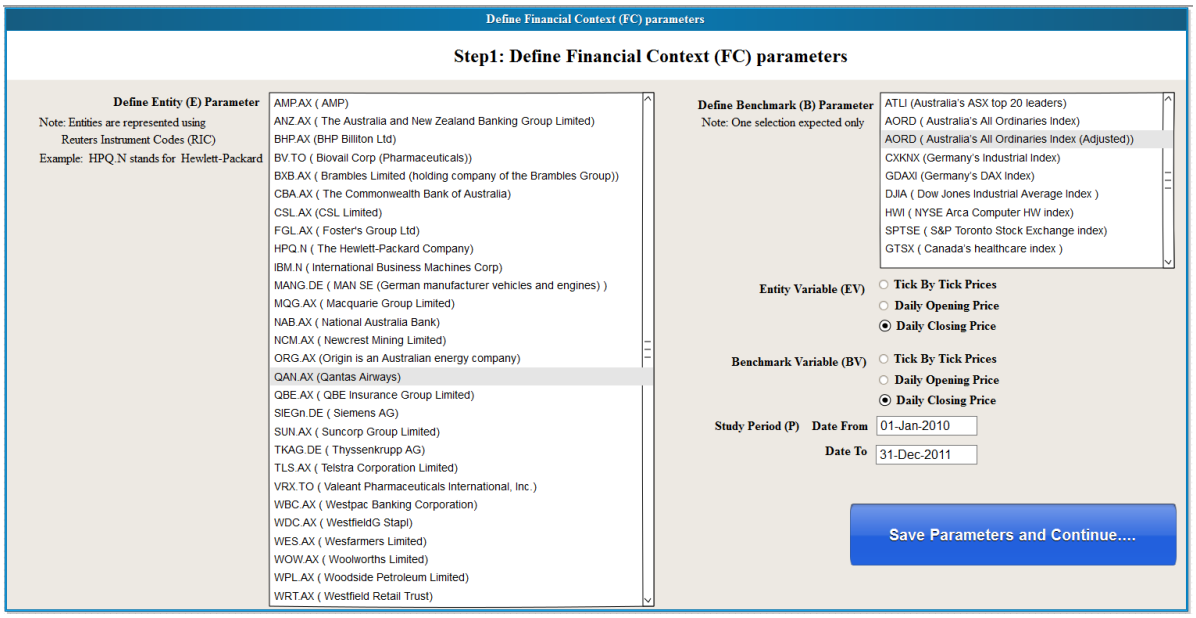


Figure B. 2 GUI to Define Financial Context (FC) parameters

The GUI shown in Figure B. 3 enables the user to define the sentiment extraction parameters for an impact study.

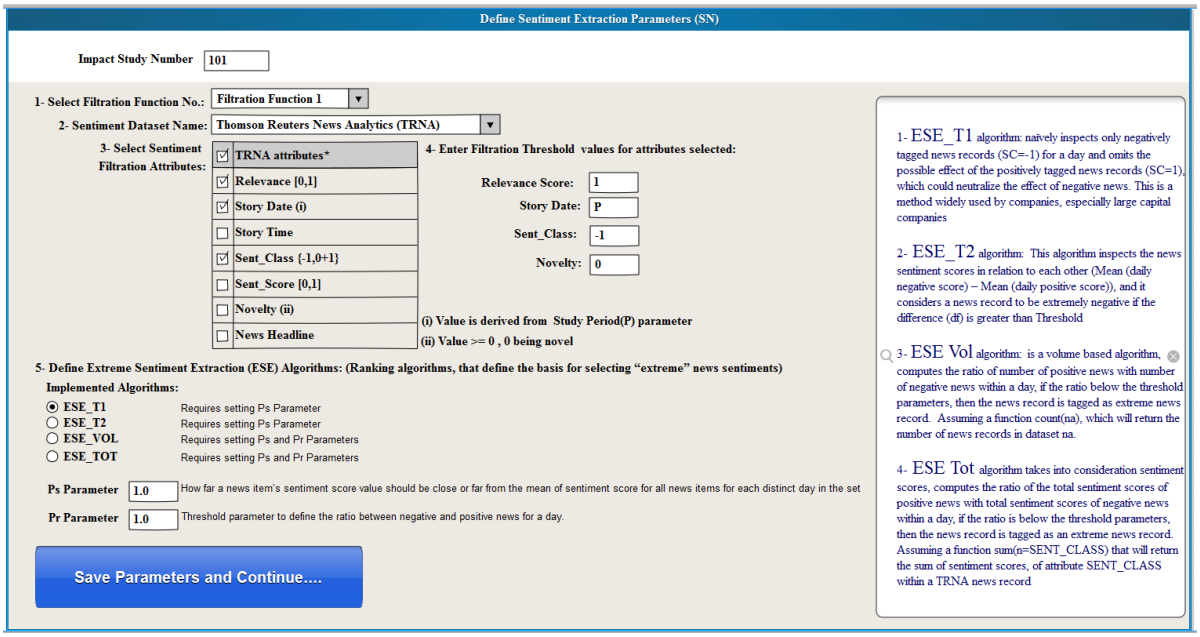


Figure B. 3 GUI to Define Sentiment Extraction (SN) Parameters

The GUI shown in Figure B. 4 enables the user to define the impact measure parameters and trigger an impact study.

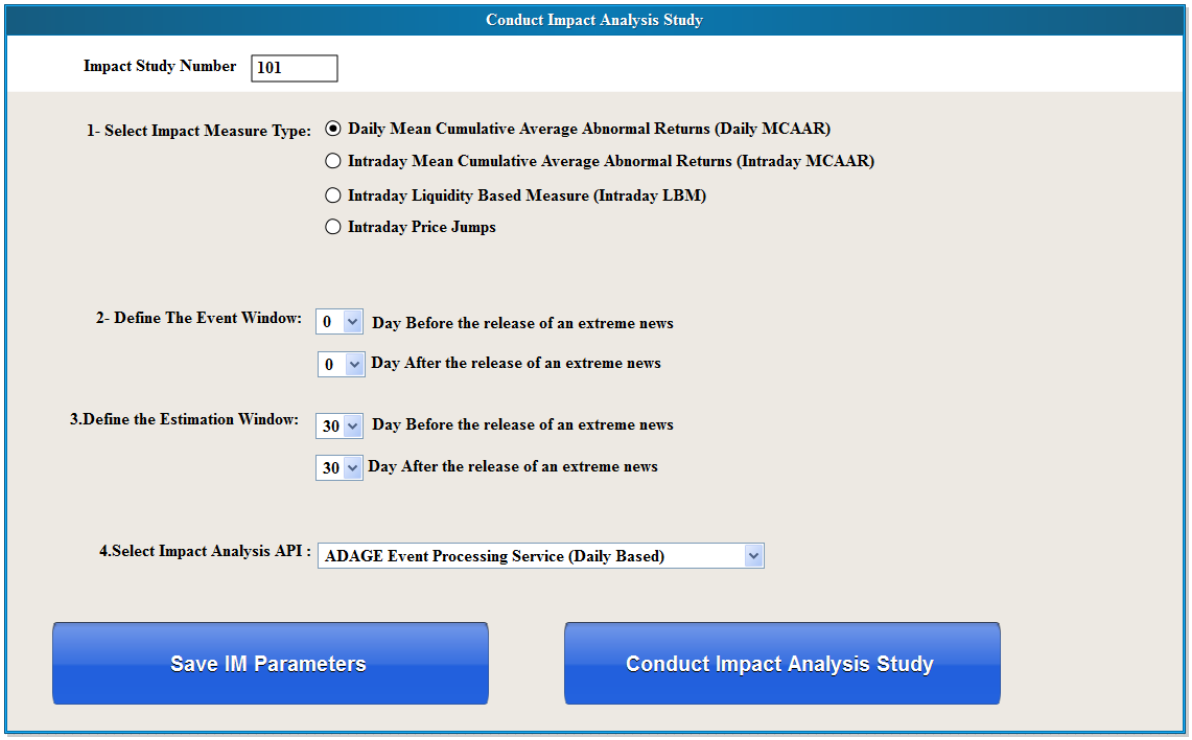


Figure B. 4 GUI to Define Impact Measure (IM) parameters and Conduct Impact Analysis Study

APPENDIX C: *EXTREME_NEWS_ALGORITHM* PSEUDO CODE

C1. Implementing ESE_T1 and ESE_T2 algorithms:

ESE_T1 and ESE_T2 algorithms make use of the following notations:

- *na* is the attribute a of news record *n.*
- *μa(S)* represents the mean of *na* in dataset (S).
- *σa(S)* is the standard deviation of *na* in dataset (S).
- *Ps* How far a news item's sentiment score value should be close or far from the mean of sentiment score for all news items for each distinct day in the set.

The algorithms pseudo code is described follows:

- ESE_T1 pseudo code:

```
ESE_T1 (Ps, TRNA_SCORES) =
V= Select all news records TRNA_SCORES table where SC= -1 // -1 Negative class
Threshold = μSS(V)+ Ps × σSS(V)
For every news record n in V do
    If SS> Threshold then
          add record n to Subset TRNA_SCORES_APPLIED
    End if
End For
Return (TRNA_SCORES_APPLIED)
```

- ESE_T2 pseudo code:

```
ESE_T2 (Ps, TRNA_SCORES) =
V= Select all news records in TRNA_SCORES table where SC= -1 // -1 Negative class
Threshold= μSS(V)+ Ps × σSS(V)
For every day d in V do
      BP = Select all news records where SC= -1 and SRD = d // -1 Negative class
      BN = Select all news records where SC= +1 and SRD = d // +1 Positive Class
      df = μSS(BN) -  μSS(BP)
```

```
      If df > Threshold then
            add subset (BN) in subset TRNA_SCORES_APPLIED
      End if
End For
Return (TRNA_SCORES_APPLIED)
```

C2. Implementing ESE_VOL, ESE_TOT and ALL_NEWS algorithms:

The prototype implements another three algorithms, two extreme sentiment extraction algorithms ESE_VOL and ESE_TOT . An additional algorithm also defined (referred to as ALL_NEWS) used for benchmarking against ESE_VOL and ESE_TOT. Both ESE_VOL and ESE_TOT use threshold parameters to rank news records in TRNA_SCORES. Table C. 1 shows the parameters thresholds used to warranty only top extreme news sentiment items are selected in each study.

Table C. 1 Extreme Sentiment Extraction (*ESE*) parameters

ESE (Pr, Ps, TRNA_SCORES)	**ESE_{VOL}**	**ESE_{TOT}**	**Description**
Pr	[0,1]	[0,1]	Threshold parameter to define the ratio between negative and positive news for a day.
Ps	[0,1]	[0,1]	How far a news item's sentiment score value should be close or far from the mean of sentiment score for all news items for each distinct day in the set

The algorithms pseudo code is described follows:

- ESE_VOL pseudo code:

```
ESE_VOL (Pr, Ps, TRNA_SCORES) =
// 1. Get Average and standard deviation of the filtered dataset TRNA_SCORES
//AvgN stores the average count of negative tagged news across the whole dataset
AvgN= μ (count (TRNA_SCORES)) where SC =-1 // -1 Negative class
//AvgP stores the average count of positive tagged news across the whole dataset
AvgP= μ (count (TRNA_SCORES)) where SC = +1 // +1 Positive Class
//StdN stores the standard deviation of the count of negative tagged news across the whole
dataset
StdN = σ (count (TRNA_SCORES)) where SC = -1
```

```
//2. Now we loop through V and get daily counts of positive and negative   tagged news records, which will be compared to average counts of the whole set stored above

V= Select all days (d) in dataset TRNA_SCORES

For every day d in V do

// dailyNCount stores the count negative tagged news records for day d

   dailyNCount(d) = Count (TRNA_SCORES)) where SRD = d and SC =- 1

// dailyPCount stores the count negative tagged news records for day d

   dailyPCount(d) = Count (TRNA_SCORES) where SRD = d and SC = +1

// Calculate the ratio of positive to negative news in day d

    RatioCount(d) = ((dailyPCount(d) ×100)/dailyNCount(d))

//If the daily ratio calculated above is less than Pr and daily count of negative news is greater than AvgN + Ps * StdN then this is a record we define as extreme

 If (RatioCount(d) ≤ Pr and

    (dailyNCount(d)> (AvgN +(Ps× StdN)))) then

       add all records of TRNA_SCORES with day d to Subset (TRNA_SCORES_APPLIED)

 End if

End For

Return (TRNA_SCORES_APPLIED)
```

- ESE_TOT pseudo code:

```
ESE_TOT (Pr, Ps, TRNA_SCORES) =

//1. Get Average and standard deviation of records in TRNA_SCORES table

//AvgN stores the mean of total of sentiment scores of negative tagged news   for all records in TRNA_SCORES table

AvgN=μ (sum (TRNA_SCORES.SENT_SCORE)) , where SC = -1 // -1 Negative class

//StdN stores the standard deviation of total of sentiment scores of negative tagged news for all records in TRNA_SCORES table

StdN = σ (sum (TRNA_SCORES.SENT_SCORE)) where SC = -1 // -1 Negative class

//2. Now we loop through V and get daily total of sentiment scores SENT_SCORE of negative and positive tagged news records, which will be compared to means of the whole set stored above

V= Select all days (d) in table TRNA_SCORES
```

```
For every day d in V do
// dailyNSum stores the sum of sentiment scores of negative tagged news records for day d
dailyNSum(d)= sum (TRNA_SCORES.SS) where SRD = d and SC = -1
// dailyPSum stores the sum of sentiment scores of positive tagged news records for day d
dailyPSum(d) = sum (TRNA_SCORES.SS) where SRD = d and SC = +1 // +1 Positive Class
// Calculate the ratio of positive to negative news in day d
  RatioTot(d) = ((dailyPSum(d) ×100)/(dailyNSum(d)))
//If the daily ratio calculated above is less than Pr and daily sum of negative news is
greater than AvgN + Ps * StdN then this is a record we define as extreme
  If (RatioTot(d)≤ Pr and
     (dailyNSum(d)> (AvgN + (Ps× StdN)))) then
        add all records of TRNA_SCORES with day d to Subset (TRNA_SCORES_APPLIED)
  End if
End For
Return (TRNA_SCORES_APPLIED)
```

- ALL_NEWS pseudo code:

```
ALL_NEWS(Filtration_Function_No)    =    Filtration_Function    (p_Filtration_Function_No,
p_Filtration_Parameters)
```