University of Strathclyde Science

TOPIC MODELLING, SENTIMENT ANALSYS AND CLASSIFICATION OF SHORT-FORM TEXT

CUSTOMER JOURNEY OF INSURANCE PURCHASES

RESEARCHER
LAZARINA STOYANOVA

CHIEF INVESTIGATOR
WILLIAM WALLACE

This dissertation was submitted in part fulfilment of requirements for the degree of MSc Information Management with Industrial Placement

DEPT. OF COMPUTER AND INFORMATION SCIENCES
UNIVERSITY OF STRATHCLYDE

AUGUST 2019

DECLARATION

This dissertation is submitted in part fulfilment of the requirements for the degree of MSc of the University of Strathclyde.

I declare that this dissertation embodies the results of my own work and that it has been composed by myself. Following normal academic conventions, I have made due acknowledgement to the work of others.

I declare that I have sought, and received, ethics approval via the Departmental Ethics Committee as appropriate to my research.

I give permission to the University of Strathclyde, Department of Computer and Information Sciences, to provide copies of the dissertation, at cost, to those who may in the future request a copy of the dissertation for private study or research.

I give permission to the University of Strathclyde, Department of Computer and Information Sciences, to place a copy of the dissertation in a publicly available archive.

Yes [✓] No [ ]

I declare that the word count for this dissertation (excluding title page, declaration, abstract, acknowledgements, table of contents, list of illustrations, references and appendices) is **21, 978**.

I confirm that I wish this to be assessed as a Type 1 2 3 4 **5** Dissertation.

Signature: Lazarina Stoyanova

Date: 19.08.2019

ABSTRACT

Upon consultation with professionals in the field of social media textual data analytics and a systematic review of literature in the field of topic modelling and sentiment analysis of short-form text, a research gap was identified, for which a prototype system was developed. The business problem faced is the lack of an automated approach to topic-sentiment extraction and classification of user-generated text based on the stage that the user is situated at in their customer journey in association with the purchase of a product or service. The following research proposes a system of tools that can extract topics and associated sentiment polarity from social media data, and subsequently allocate user-generated text in pre-defined classes that correspond with stages of the customer journey.

The research involved experimental procedures in the field of sentiment classification, topic modelling and text classification. To evaluate the models' performance a survey was distributed, which engaged a total of 58 respondents to perform the same tasks that the algorithms were given. The technical and human-agent experiment results were compared with the aim of evaluating the ability of an automated approach to solve this business challenge in a timely and efficient manner, which would emphasise the organisational benefits of cost-cutting and intelligent decision-making, which could be achieved following the implementation of the system. Considering the scope of the research project, the data used was extracted from social media websites Facebook and Twitter, and thus lacked labels, hindering the application of supervised learning for the classification task. Nonetheless, unsupervised and semi-supervised approaches were implemented, with the script for supervised model being annexed to support the work of other researchers.

The conceptualised system of algorithms has measurable benefits to organisations and has been approved for implementation as part of the initial stages of a strategic project in the University of Strathclyde. The research presents exciting opportunities for future research, as well as actionable recommendations and implications for both text analytics professionals, business owners and academics.

ACKNOWLEDGEMENTS

I would like to express my sincerest gratitude to the project's chief investigator - William Wallace, who offered valuable guidance throughout the duration of the project. His experience as a knowledge exchange fellow greatly complimented my background in business, resulting in a collective thought process that is centred around making an impact and applying technical knowledge in a real-world context through process automation and implementation of machine learning, with very many interesting discussions along the way.

I would also like to thank my all colleagues at the University of Strathclyde, who offered their support for me and expressed encouragements for all my ideas. I am grateful and humbled to be part of the Collaboration Services team and look forward to implementing the acquired knowledge to better the University's processes.

Last, but not least, my sincerest gratitude goes to my family, especially my mother, who acted as a friend, motivational speaker, therapist and comedian, whenever necessary, offering her unconditional support throughout the pursuit of my degree. Without you this all would not have been possible… благодаря ти, мамо.

TABLE OF CONTENTS

Declaration iii

Abstract iv

Acknowledgements v

Table of contents vi

List of Figures xi

List of Equations xii

List of Tables xiii

List of Abbreviations xiv

Glossary xv

1. Chapter I: Introduction 16

1.1 Introduction 16

1.2 Background to study 16

1.3 Statement of Problem 17

1.4 Purpose of Study 18

1.5 Research Questions 18

1.6 Research Design 18

1.7 Definition of Key Terminology 19

1.8 Significance of Study 19

1.9 Contributions 20

1.10 Limitations, Research Context and Scope 20

1.11 Organisation of Study 21

2. Chapter II: Literature Review 22

2.1 Chapter Overview 22

2.2 Methodology 22

2.3 Topic Modelling 25

2.3.1 Concept Overview....25
2.3.2 LDA-based (latent Dirichlet allocation): Application and Limitations....25
2.3.3 Other approaches: Applications and Limitations....27
2.4 Sentiment Analysis....28
2.4.1 Concept Overview....28
2.4.2 Primary Methods: Applications and Limitations....29
2.5 Challenges of Short-form text Topic Modelling and Sentiment Analysis....31
2.6 Topic Modelling and Sentiment Analysis of Short-Form text....32
2.7 Identification of a Research Gap....36
2.8 Conceptual Model....37
2.9 Conclusion....38
3. Chapter III: Methodology....39
3.1 Introduction....39
3.1.1 Research Questions and Hypotheses....39
3.1.2 Deliverables....40
3.1.3 Chapter Structure....40
3.2 Research Overview....40
3.2.1 Research Philosophy....40
3.2.2 Research Paradigm....41
3.2.3 Research Strategy....42
3.3 Techniques and Procedures....43
3.3.1 System Requirements....43
3.3.2 System Development....44
3.3.2.1 Data mining and associated procedures....44
3.3.2.2 Data pre-processing....45
3.3.2.3 Data exploration and Feature Extraction....49
3.3.2.4 Sentiment Analysis Experiment Design....52

3.3.2.5 Topic Modelling Experiment Design....53
3.3.2.6 Text Classification Experiment Design....54
3.3.3 System Optimisation and Testing....55
3.3.3.1 Comparative Performance Evaluation Procedures....55
3.3.3.2 Potential for System Optimisation and Parameter Fine Tuning....56
3.3.3.3 Comparative Performance Evaluation through Human-agents....56
3.3.3.4 Prototype Development....57
3.4 Limitations of Research....58
3.5 Ethical Considerations....58
3.6 Evaluation of Academic Rigour....59
3.6.1 Replicability....59
3.6.2 Reliability and Triangulation....60
3.6.3 Validity and Generalisation....60
3.7 Conclusion....60
4. Chapter IV: Analysis....61
4.1 Introduction....61
4.2 Sentiment Analysis Experiment Results....62
4.2.1 Presentation of Results....62
4.2.2 Comparative Analysis of Automatic and manual sentiment Classification....63
4.2.3 Discussion of results....65
4.3 Topic Modelling Model Evaluation....65
4.3.1 Presentation of Results....65
4.3.2 Human-Agent Performance Evaluation Results....68
4.3.3 Discussion....70
4.4 Text Classification Model Evaluation....71
4.4.1 Presentation of Results....71
4.4.2 Human-Agent Performance Evaluation Results....73

4.4.3 Discussion 75
4.5 Conclusion 76
5. Chapter V: Conclusion and Recommendations 77
5.1 Introduction 77
5.2 Recap of Problem Statement 77
5.3 Key Findings and Associated Conclusions 78
5.4 Recommendations and Implications for key Stakeholders 79
5.4.1 Presentation of Demo System 79
5.4.2 Process Automation 79
5.4.3 Working with real-time, unlabelled, short-form data 80
5.5 Future Research Opportunities 80
5.5.1 Application in Learning Analytics and Education Enhancement for University of Strathclyde 80
5.5.2 Application in Social media analytics for Hospitality and Tourism Industry 82
5.5.3 Academic Research Experimental Opportunities for System Enhancement 82
5.6 Personal Reflection Statement 83
5.7 Conclusion 83
References 84
Appendices 103
A MyCustomerLens Company Profile 103
B The Insurance Industry in 2019: Market Overview 104
C Systematic Literature Review Methodology Process 105
D Data Extraction Procedure: Limitations, Stats and Queries 107
E Ethical Approval Confirmation 109
F Research Protocol 110
G Electronic Consent Form 113
H Survey Response Data 120
I Location Map of Study Participants 155

J LDA Topic Model Term Probability Demonstration on Selected Texts 156

K Annex Documentation: Supporting Python Code, Extracted Data and Survey Data 158

LIST OF FIGURES

Figure 2.2-1 Summary of Search Strategy, using the PRISMA flow methodology (see Moher et al., 2009) 23

Figure 2.2-2 Distribution of Studies (returned from database search alone) per Journal and Year. 24

Figure 2.8-1 Conceptual model of proposed research (methodological) 38

Figure 3.3-1 Data Cleaning Pipeline 46

Figure 3.3-2 Stopwords Count in Individual Text Entries 47

Figure 3.3-3 Feature extraction pipeline 49

Figure 3.3-4 Data entries individual count of words per entry (left) and characters per entry (right) 50

Figure 3.3-5 Most Frequent words in the Dataset, represented in a Word Cloud Format 50

Figure 3.3-6 All text classification models, developed as part of the current research, arranged by approach type 55

Figure 4.1-1 Age Distribution of Survey Participants 61

Figure 4.2-1 Histogram of Sentiment Polarity, extracted from Textblob sentiment classification 62

Figure 4.2-2 Naive Bayes sentiment classification result 63

Figure 4.3-1 LDA topic models (Word Cloud Representation) 67

Figure 5.4-1 Demo Mobile App Functionality Prototype 79

LIST OF EQUATIONS

Equation 3.3-1 Vectorisation Methodology 52

Equation 3.3-2 Naive Bayes theorem.....53

Equation 3.3-3 Matrix decomposition of LDA and LSA topic modelling techniques (Bergamaschi and Po, 2014).....54

LIST OF TABLES

Table 2.4-1 Comparison of subjectivity detection semantic methods (Overview) (adapted from Chaturvedi et al., 2018) 30

Table 2.6-1 Comparison of Deep Learning methodologies (adapted from Dohaiha et al., 2018) 33

Table 2.6-2 Overview of Extracted Models, Methods and Applications for Simultaneous Topic Modelling and Sentiment Analysis of Short-form text 35

Table 3.2-1 Summary of Research Ontology (adapted from Saunders et al., 2016). 43

Table 3.3-1 System Requirements Catalogue Brief (adapted from IEEE Computer Society, 1999; 2009) 43

Table 3.3-2 Most common and Least Common Word List 48

Table 3.3-3 Methodological Desicions concerning survey experiment with human participants (a summary) 56

Table 4.2-1 Word cloud with negative (left) and positive (right) sentiment polarity, extracted from textblob classification 63

Table 4.2-2 Comparative Analysis of results from manual and automated (lexicon-based) sentiment classification on selected user-generated texts 64

Table 4.3-1 LSA topic modelling results 66

Table 4.3-2 Evaluation Metrics for LDA performance 67

Table 4.3-3 Manually generated topic models by study participants for selected texts 68

Table 4.4-1 K-Means clustering text classification performance 71

Table 4.4-2 Word-level Dictionary-based Text Classification with Naive Bayes Results 72

Table 4.4-3 Results from manually generated by study participants text classification on selected texts 73

Table C-5.7-1 Systematic Review Worksheet, based on the PRISMA methodology (adapted from Moher et al., 2009) 105

Table D-5.7-2 Facebook Data Extraction Procedure, Key Stats and Limitations 107

Table D-5.7-3 Twitter Data Key Stats and Limitations 108

LIST OF ABBREVIATIONS

API – Application programming interface

C2C – Customer to Consumer

B2C – Business to Consumer

GDPR – General Data Protection Regulation

GUI – Graphical User Interface

ESRC – Economic and Social Research Council

SVM – Support Vector Machines

IDF – Inverse Document Frequency

TF – Term Frequency

NLP – Natural Language Processing

SVD – Singular Value Decomposition

IoT - Internet of Things

NLTK - Natural Language Toolkit

ILE - Interactive Learning Environment

PR – Personal Relations

GLOSSARY

Sklearn - Scikit-learn (formerly scikits.learn) is a free software machine learning Python library, which features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

Textblob - a Python library for processing textual data, which provides a simple API for common NLP tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

Word2vec - a group of related models that are used to produce word embeddings from text

Customer journey – a theory, used predominantly in marketing practice that represents the process that a customer, purchasing a product or services undertakes; consists of five stages: expectation/awareness, consideration, purchase, retention, advocacy

1. CHAPTER I: INTRODUCTION

1.1 Introduction

The following research is concerned with addressing the business problem of short-form text handling automation. Specifically, a design of a system will be provided following the execution of experiments in three areas of Natural Language Processing (NLP): Topic Modelling, Sentiment Analysis and Text Classification, with the latter being done in a way that indicates to the system's user (typically a business organisation) the stage of the customer journey (typically associated with a purchase of a product or service) that the author of the text is at.

The Introduction chapter will present the background and context of this research and the research questions that will be tackled. Other elements of design will also be discussed, as well as the contributions of this piece being affirmed prior to delving in deeper into the subject matter.

1.2 Background to study

According to market research of the digital market, the number of internet users between January 2018 and January 2019 has grown by 9.1% or otherwise 367 million reaching a total of 4.388 billion internet users, with active social media users and mobile social media users following a similar trend of growth for the same period, with 9% and 10%, respectively (Kemp, 2019: 8). Social media platforms continue growing in popularity, such as Facebook or Instagram, who have in the past year gained 37 and 38 million new active followers, respectively, which translates to 1.7% and 4.4% of the corresponding user base of these platforms (Kemp, 2019: 82). Users of such social media collectively post online vast amounts of data, which are considered by business organisations and market researchers as sources of market research data, available to the public. However, relevant insight is difficult to find as data is often considered a chaotic cluster of various information formats (Ritter et al., 2011; Linoff and Berry, 2011) or can offer minimal insight to marketing and business strategists.

Partially influenced by these problems, the field of NLP has been intensively developed in recent years, its aim being to train algorithms to decode natural language and speech data into meaningful semantic insights through processing, analysis and synthesis, bridging the gaps in communication between humans and machines (Nadkarni et al., 2011). Applications are thus being developed, whose aim is to understand sociological constructs through computer science (Wang et al., 2007), and translate trending social insights to marketing specialists. Ultimately, the goal of research in intelligent social media analytics software is to measure consumer response to stimuli and events, and report insights that can improve organisations' competitive advantage (Amaravadi et al., 1995; Chen et al., 2012) as they adapt their B2C communication and content

dynamically (Nakatani and Chuang, 2011). Consequently, a common NLP research problem is the extraction of sentiment from text, classifying an expressed opinion as positive, negative or neutral, which is used by analysts to better understand societal response to trends and pressing issues (Pang and Lee, 2008; Fan and Gordon, 2014; Liu and Zhang, 2012). Opinion extraction from social media data offers challenges, such as language ambiguity or expression of mixed semantic attributes (Liu and Zhang, 2012), as well as such insight being arguably challenging for marketers to translate into sales or purchase intent due to a lacking context of the opinion (Omand et al., 2012). The current research will approach the problem of sentiment analysis and topic extraction of social media data, while simultaneously addresses the needs of marketing specialists through classifying consumer-generated text into stages of the customer journey.

1.3 **Statement of Problem**

Due to the potential applications of sentiment analysis and topic modelling instruments in the context of understanding consumer behaviour and informing business decisions, research in the areas has been intense in the past few years. However, user-generated social media textual data offers multiple challenges for the development of algorithms, such as data sparsity (Chen et al., 2011; Rao et al., 2016; Ittoo et al., 2016), lack of structure (Oza and Naik, 2016; Curiskis et al., 2019) and lack of annotation (Curiskis et al., 2019) to name a few, which will be further expanded on in Section 2.5 of Chapter II. Research has thus progressed from surface-level traditional machine learning modelling to deep learning state-of-art methodologies, that are more adaptive to unstructured and unlabelled data, and can automatically extract features and rich data representations (Araque et al., 2017). As a result topic-aware sentiment analysis has become more accessible as a research discipline, with various scholars proposing models that can be used for social media data (see Rao et al., 2016; Diamantini et al., 2019; Li et al., 2019; Ali et al., 2019; Huang et al., 2017; Zhang et al., 2016; Dong et al., 2018; Fu et al., 2018; Xiong et al., 2018; Ren et al., 2016; Farhadloo et al., 2016). Although such models have been tested in various domains, such as financial markets (Nassirtoussi et al., 2014), politics (Lozano et al., 2017), and retail (Ibrahim and Wang, 2019) (see more in Section 2.7, Chapter II), it is recognised that few studies have examined topic modelling and sentiment analysis as means of supporting marketing decision-making. Specifically, as Chapter II: Literature Review will demonstrate, few recent studies have addressed the knowledge gap of applying a classification algorithm as a subsequent step to topic-sentiment models. Considering also the above demonstrated gap of research that supports the function of marketing personnel, the current study aims to create a system of tools that can extract topics and associated sentiment polarity from social media data, and subsequently allocate user-generated text in pre-defined classes that correspond with the stages of a purchase customer journey.

1.4 Purpose of Study

The purpose of the study will not be to propose new development of algorithms for topic modelling, sentiment analysis or text classification, but to examine the performance of existing algorithms in a collective system, whose aim is solving a real-life business problem, as explained above. Most importantly, a combination of existing techniques will be sought that solves the challenges of working with short-form text in the most time-efficient manner. Finally, performance of compared systems will be evaluated on the basis of technical performance, ease of application, as well as proximity to human agent performance on the same problem, which collectively will act as a determinant of system quality. Ease of application is especially emphasised considering that one of the key reasons for the creation of this system being the desire to automate previously manual processes in ways that can be applied directly in small marketing and business organisations, as well as are scalable for use in larger corporate entities.

1.5 Research Questions

Motivated by the problems identified in Sections 1.2 and 1.3, the following research will address the research questions listed below:

- *Can a library-generated sentiment classifier replace a manual sentiment classification process efficiently?*
- *Which topic modelling technique can be considered most efficient for handling of short-form, user-generated social media textual data, with experimentation comparatively evaluating the performance of LSA and LDA for topic coherence and similarity with topics generated by humans on a small sample of the data?*
- *Which classification technique can be efficiently applied to a web-extracted dataset with user-generated text to categorise the data entries into five categories that correspond with the user journey?*

1.6 Research Design

The design of the proposed research is therefore quantitative, with all associated activities being conducted in a scientific and experimental manner that suggests that all derived insight is supported by empirical data. Whenever such is not available, qualitative interpretation is incorporated. Based on Saunders et al.'s (2016) research onion ontology, the research philosophy is positivism with a deductive approach, and a cross-sectional time horizon. A systematic literature review in the field of topic modelling and sentiment analysis of short-form text will reveal the algorithms that are most suited for inclusion in the experimentation process. Subsequently, the performance of each combination of algorithms will be assessed

The holistic system development methodology followed is derived from Géron's (2017) machine learning project checklist; however, appears in a variety of texts, e.g. Chollet (2018), Nielsen (2015), Goodfellow et al. (2016), Russell and Norvig (2016) and Berry and Linoff (2004), where the project's milestones involving problem framing, obtaining data, data exploration, data preparation, short-listing of promising models, system fine-tuning and solution presentation. Considering Fernandez-Lozano et al.'s (2016) critical evaluation of this traditional experiment design template in computational intelligence, one adjustment is made. External cross-valuation is introduced in the learning stage, in the current research done by human-agent evaluation. This evaluation will be made available to the study participants in the form of an online survey. The details regarding the use of this instrument, as well as its protocol and measures are available in Section 3.3.3.3, in the Methodology chapter.

Data sourcing will be done through accessing publicly available social media user-generated texts from the platforms Facebook and Twitter, with the rationale and methodology applied for data access and pre-processing being explained in detail in Section 3.3.2.1, in the Methodology chapter. A demo presentation of the final solution will also be made available as part of the deliverables of this research project.

1.7 Definition of Key Terminology

Considering the business challenge that is being addressed with the system development, the term 'customer journey' requires further clarification. The customer journey concept is a key aspect of marketing theory (Rawson et al., 2013), with multiple interpretations available, e.g. a user story (Stickdorn and Schneider, 2010), or the repeated interactions between a service provider and the consumer (Sangiorgi, 2011). Holistically, the concept implies that each customer of any organisation goes through five stages as part of their purchasing process: expectation/awareness, consideration, purchase, retention, advocacy (Følstad and Kvale, 2018; Voorhees et al., 2017; Lemon and Verhoef, 2016), which will be used as classes (categories) for the machine learning classifier.

1.8 Significance of Study

From an organisational standpoint, the ability to relate sentiment to given topics enables informed planning of business operational goals, with the capacity to prioritise areas, identified as problematic. Relating the topic-aware sentiment analysis to stages of the customer journey enables improvements in targeted responsiveness of the organisation and as a result – improved communication with consumers and feel for the market. Such technology can empower organisations to monitor consumers and their responses to stimuli intelligently, whilst simultaneously taking a proactive response to identifying the topics, which interest consumers at

various stages of their customer journey and tracking the associated sentiment consumers have with such topics. Arguably, such information except from a strategic standpoint, has value from a marketing standpoint as well, namely for aspects of digital marketing such as the business' content strategy. Understanding the topics that are relevant to consumers at each stage of the customer journey enables organisations to target market micro-segments with marketing communication or promotional activities. In addition, this increases the likelihood of immediate, personalised responses, which has the potential of improving companies' relationship marketing efforts, which as a result can improve customer retention. Moreover, being able to capture sentiment associated with individual topics in the journey stages can lead to identification and understanding of process 'leaks', otherwise stages that can be associated with loss of consumers. Such knowledge can be used for strategic process improvement with the aim of retaining consumers.

From a research standpoint, the current piece advances literature by demonstrating the potential in combining existing machine learning algorithms from different disciplines in an effort to solve a real-life business problem. Specifically, the research identifies in a scientific manner the superiority of a number of techniques compared to others that serve the same purpose, which knowledge can be utilised by other researchers as a starting point in their own system development in the field of topic modelling and sentiment analysis of short-form text.

1.9 Contributions

The study will make contributions to the development of industry practice, demonstrating cost- and time-efficient ways of implementing machine learning for marketing process automation, as well as academic contributions, which will be deriver as a result from the experimental activities that involve the comparative testing and performance evaluation of models against one another, as well as against human-agents (study participants).

1.10 Limitations, Research Context and Scope

Considering the work with user-generated social media data, a key limitation is the quality and availability of data. To elaborate, the importance to training data for a machine learning algorithm of any type is pivotal for its performance, as recognised by a variety of scholars (see Géron, 2017; Chollet, 2018; Nielsen, 2015; Goodfellow et al., 2016; Russell and Norvig, 2016). Although more advanced methods can be used to address this challenge, e.g. deep learning methods, the problem with imbalanced (Chawla et al., 2004) or insufficient (Aggarwal and Zhai, 2012) data is presented in the training stage.

In order to potentially capture data from various stages of the customer journey, a specific industry context should be examined. Following consultations with company executives in the industry of marketing analytics that specialise in textual data (see Appendix A), the insurance industry was chosen as a suitable cohort, key information for which is attached as Appendix B. The rationale for choosing this industry amongst others was the long duration of the customer journey that characterises B2C relationships in the field, which is considered beneficial for addressing the above limitation of data insufficiency or imbalance.

A key assumption that underpins the choice made is (1) the duration of a mandated relationship between an organisation and a consumer (e.g. through an insurance policy), in combination with (2) the increased psychological investment of the consumer in the process of decision-making, which is affirmed by research suggesting that insurance is a high-involvement[1], self-concern purchase (see Lin and Chen, 2006; Mittal, 1989; Kim and Sung, 2009). Collectively these assumptions are considered to increase the likelihood of social interactions in the digital space that concern various stages of the customer's journey.

Regardless of the specific nature of the context, the research is argued to have external generalisability from a system perspective, with the industry being determined by the data the system is trained on. Further details of how academic rigour is ensured are available in Section 3.6 in the Methodology chapter.

1.11 Organisation of Study

The following Chapter II: Literature Review will present a systematic synthesis of relevant literature in the form of a literature review, where the knowledge gap this study aims to fulfil will be contextualised.

[1] High-involvement purchases are found to absorb more consumer time in the stages of information seeking and consideration, resulting in more time and more money being spent for such a purchase (Clarke and Belk, 1979). Although no formal definition exists, high-involvement purchases are made by conscious consumers, who for a variety of reasons consider the outcome of the purchase to be of critical importance to their life (Park et al., 2007).

Chapter III: Methodology will situate the study within a precise methodological tradition, explaining the rationale for relevant decision-making, associated with all aspects of the design and procedures that were part of the study. Chapter IV: Analysis is where the findings will be presented and results - critically analysed, considering the study's research questions, literature review, and conceptual framework. The final chapter (Chapter V: Conclusion and Recommendations) is where the outcome of the study will be discussed, specifically the patterns, ambiguities or inconsistencies of the findings, as well as personal reflection statements, concerning the research process and future research opportunities that stem for the current study.

2. CHAPTER II: LITERATURE REVIEW

2.1 **Chapter Overview**

The following chapter will provide a detailed account of scholarly work that has been previously published in the examined areas of topic modelling and sentiment analysis. The aim of the review is to systematically examine recent publications in the area and familiarise the reader with recent developments. A further objective is to approach knowledge from a critical stance, as well as demonstrate the gaps in knowledge that will be addressed by the current research.

An overview of the methodology used for completing the review will be provided, after which topic modelling and sentiment analysis literature will be examined in separate sections. The most relevant studies to the current research are synthesised and critically analysed in Sections 2.6 and 2.7, where knowledge gaps will be discussed. In Section 2.8 is attached a conceptual model, which will detail the theoretical and methodological bases for development of the study and analysis of findings, following which the chapter will be concluded with a brief overview.

2.2 **Methodology**

The adopted literature review methodology is systematic. Reviews are considered systematic if they adhere to a methodological approach that is (1) explicit in terms of defining the procedures followed in the process of conducting the review, (2) comprehensive in scope regarding the inclusion of all relevant material on the given topic, and as a result is (3) reproducible by others, following the same approach in reviewing the topic (Fink, 2005; Jesson et al., 2011; Booth et al., 2016; Hart, 2018). Key characteristics of such reviews are also transparency regarding the approach, rationale and decisions made by the researcher (Rousseau et al., 2008). Considering critics of the traditional graduate student thesis approach to conducting a literature review, the following examination, although scope-limited, follows the methodological steps of a stand-alone, systematic literature review, closely mirroring methodologies of doctoral theses, namely that selected studies meet rigorous characteristics for the independent and dependent variables (Okoli

and Schabram, 2010). The following eight procedures have been conducted: purpose identification, development of protocol, search for literature, practical screen, quality appraisal, data extraction, synthesis of studies and writing of the review, which can be loosely grouped into four stages: planning, selection, extraction and execution (Okoli and Schabram, 2010; Jesson et al., 2011).

Figure 2.2-1 Summary of Search Strategy, using the PRISMA flow methodology (see Moher et al., 2009)

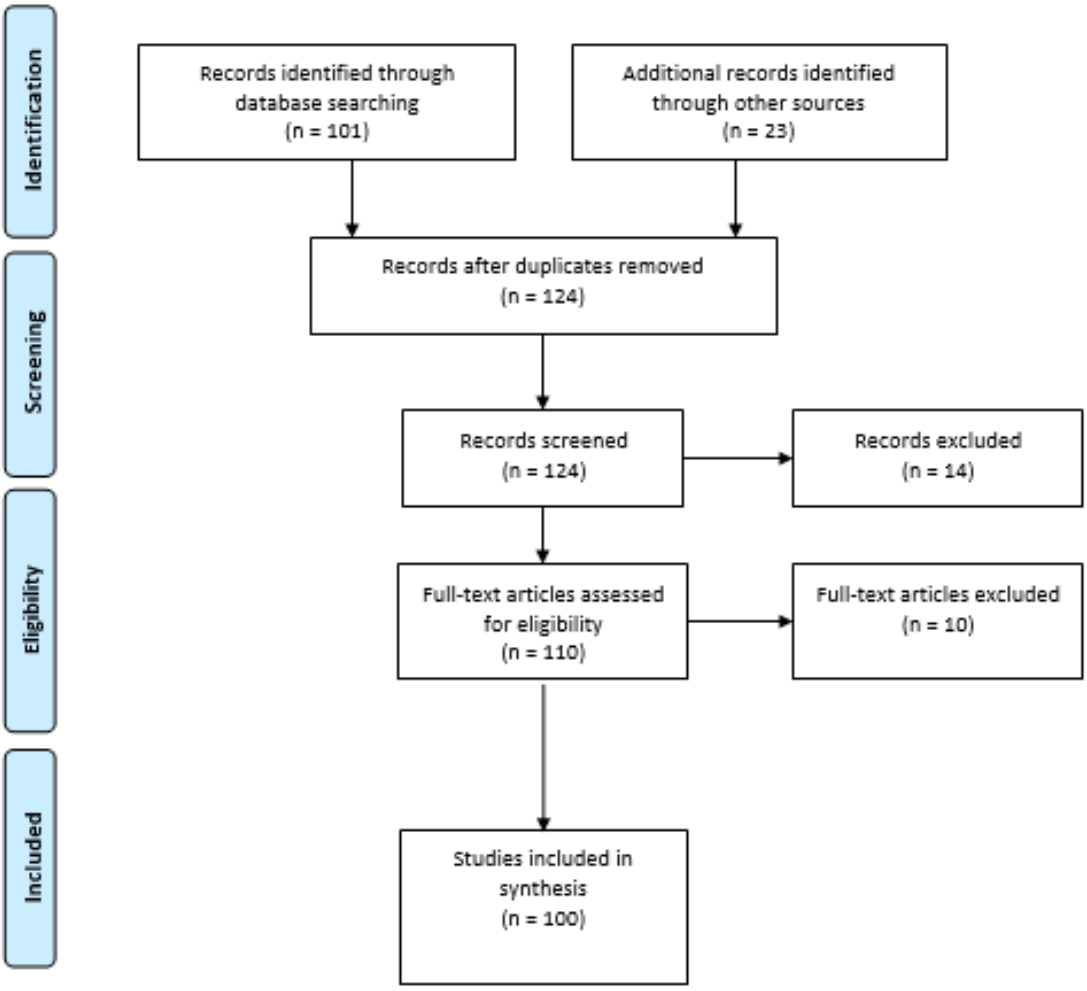
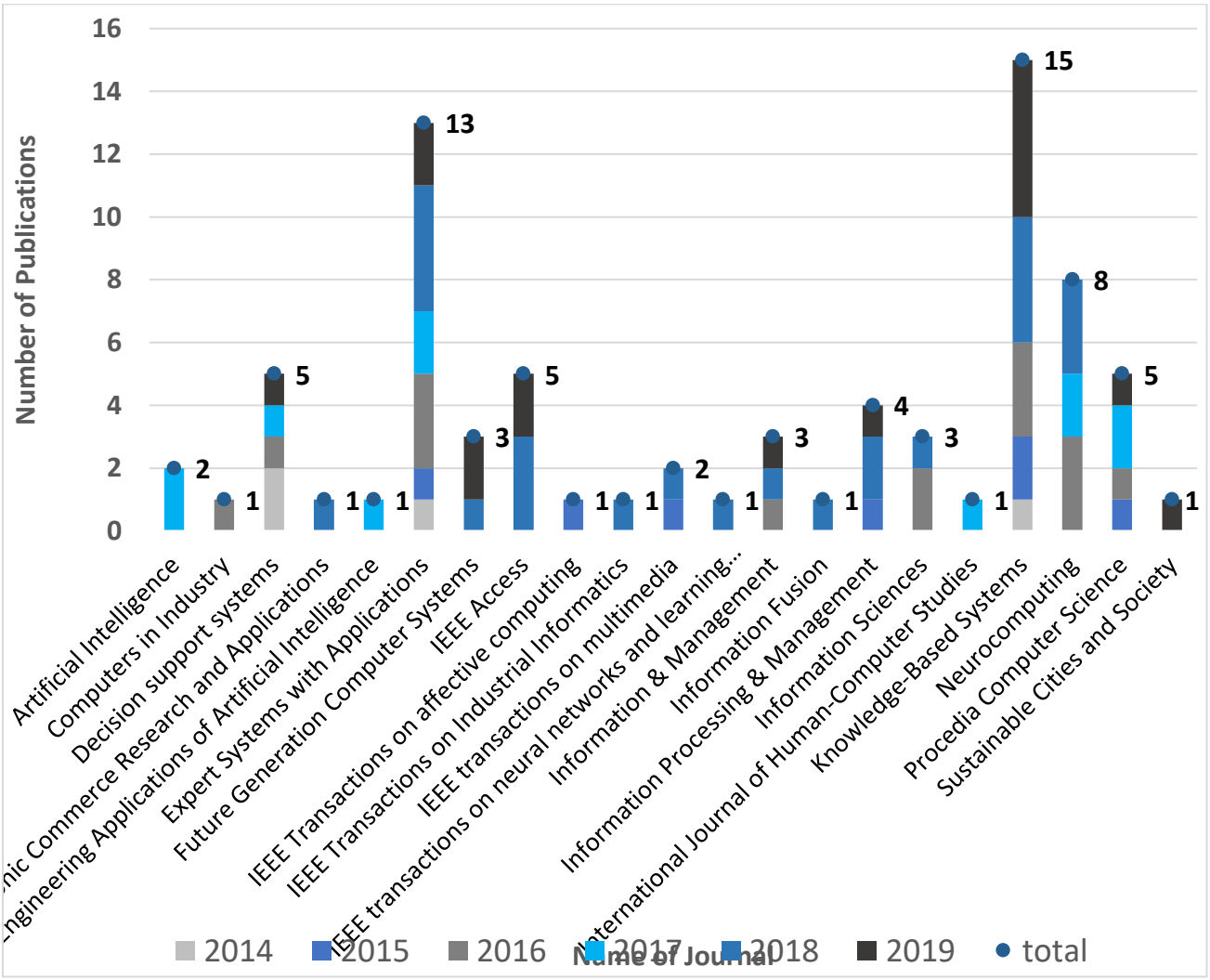


Figure 2.2-1 (above) illustrates a summary of the applied search strategy, which mirrored the PRISMA methodology for systematic literature reviews, developed by Moher et al. (2009). To ensure the reproducibility of the review, details of conducted searches are attached as Appendix C. Several decisions are to be justified. Firstly, database choice was made on the basis of optimal search, with electronic databases chosen for efficiency. Elsevier and IEEE Xplore were chosen as examples of industry and context-specific databases, respectively. Considering the more general nature of publications on the Emerald Insight database, as well as its poor performance

in retrieving relevant papers, subsequent handsearching was performed in Google Scholar. Secondly, inclusion and exclusion criteria were applied, most notable among the latter being journal quality, measured by the ABS Academic Journal Guide of 2018 (CABS, 2019) or the Web of Science Journal Impact Factor Index of 2017 (Clarivate Analytics, 2019). By setting these criteria, the scope of the synthesis was placed on original and well executed research papers in highly regarded journals, thus enabling a coherent examination of the discussed topic and the knowledge gaps, which can be addressed.

Finally, an overview is provided of the distribution of studies per journal and year, from those returned by the database search alone, which were subsequently selected for analysis following application of the exclusion criteria illustrated above (Figure 2.2-2, below). This demonstrates the leading journals, as well as the trend in publications in the field, namely the majority of relevant papers being published in 2018 and 2019, indicating an upward trend in academic popularity, available knowledge and researcher interest in the field. The following sections will present the literature analysis, organised by topics, as indicated earlier in Section 2.1.

Figure 2.2-2 Distribution of Studies (returned from database search alone) per Journal and Year. [2]



2.3 **Topic Modelling**

2.3.1 Concept Overview

Topic modelling is a text processing technique, which is aimed at overcoming information overload by seeking out and demonstrating patterns in textual data, identified as the topics (Blei et al., 2003). This enables an improved user experience, with users being equipped with the ability to navigate quickly through a corpus of text or a collection, guided by identified topics (Blei and Lafferty, 2007). Primarily topic modelling is performed with unsupervised learning algorithms, the output of which is a summary overview of the discovered themes (Lee et al., 2017). Topic detection can be performed in either online of offline mode, with the former aiming to discover

[2] *Note: Additional Studies were included as a result of handsearching and reference list searching*

dynamic topics overtime as they appear and the latter being retrospective, considering documents in the corpus as a batch, detecting topics one at a time (Chen, Guo et al., 2017). There are, according to Dang et al.'s (2016) literature review, four main approaches to topic detection and modelling: keyboard-based approach, probabilistic topic modelling, Aging theory, and graph-based approaches. Other scholars consider categories being best defined by techniques used for topic identification, such as clustering, classification or probabilistic techniques (Cigarrán et al., 2016).

2.3.2 LDA-based (latent Dirichlet allocation): Application and Limitations

LDA (Latent Dirichlet Allocation) is a Bayesian hierarchical probabilistic generative model for collection of discrete data and it operates based on an exchangeability assumption for words and topics in the document (Blei et al., 2003). In this method, documents are modelled as discrete distributions over topics, and later topics are regarded as discrete distributions over the terms in the documents (Wang et al. 2018). The original LDA method uses a variational expectation maximization (VEM) algorithm to infer topics for LDA (Blei et al., 2003), but later stochastic sampling inference based on Gibbs sampling was introduced, which demonstrated improved performance in experiments and has since been used more frequently as part of models (Wang et al. 2018). Blei et al. (2003), who first introduced LDA demonstrate its superiority against the probabilistic LSI model. LSI (Latent Semantic Indexing) contrastingly uses linear algebra and bag-of-words representations for extracting words with similar meanings (Kintsch et al., 2007). Its limitations involve its ability to scale due to the linearity of the technique it is based on, however pLSI, the probabilistic variant of LSI, solves this challenge by using a statistical foundation instead and working with a generative data model (Uys et al., 2008; Onan et al., 2016). Nonetheless, LDA was most commonly listed as part of models amongst all reviewed techniques and is considered of value for strategic business optimisation. For example, Wang et al.'s (2018) study demonstrates the value of the methodology as means of improving a company's competitive advantage by extracting information from user online reviews, and subsequently classifying topics according to sentiment. Although Wang et al.'s (2018) paper demonstrates meaningful findings and a system easily utilisable by managers, it fails to provide comparative analysis that can potentially demonstrate the superiority of the proposed model architecture. Topic modelling using LDA has been used also to characterise personality traits of users, based on their online text publications (Liu et al., 2016). Notable is also the study of Bastani et al. (2019), where LDA-based topic modelling is used to analyse consumer complaints in a consumer financial protection bureau. As part of these models, predetermined labels are used for classification, which improves the efficiency of the complaint handling department through task automation.

Although efficient and frequently used in scholarly research, the model is criticised for its assumption of document exchangeability, which can be restrictive in contexts where topics evolve overtime (Uys et al., 2008). Additionally, LDA-based models are criticised for commonly neglecting co-occurrence relations across the documents analysed, which results in detection of incomplete information and an inability to discover latent co-occurrence relations via the context or other bridge terms, which subsequently prevents topics that are important but rare from being detected (Zhang et al., 2016). Hybrid approaches have been proposed to address these limitations (Zhang et al., 2016), however they perform sub-optimally on short-form text, which brings to question their efficiency in noisy, unstructured social media data. This criticism is also shared in the analysis of Curiskis et al.'s (2019) study, where the authors propose a model specifically tailored for online social networks topic modelling, demonstrating that even shallow machine learning clustering techniques applied to neural embedding feature representations deliver more efficient performance as compared to LDA. Models, who learn vector representations of words and hidden topics are justified to have a more effective classification performance on short-form text (Zhang and Zhong, 2016). Similarly, Yu and Qiu (2018) propose a hybrid model, where the user-LDA topic model is extended with the Dirichlet multinomial mixture and a word vector tool, resulting in optimal performance, when compared to other hybrid models or the LDA model alone on microblog textual data. Similarly, Yu et al. (2019) apply a conceptually similar approach to Twitter data, namely the hierarchical latent Dirichlet allocation (hLDA), which aims to automatically mine the hierarchical dimension of tweets' topics by using word2vec (i.e. a vector representations technique) to extract semantic relationships of words in the data to obtain a more effective dimension. Hajjem and Latiri (2017) further criticise the LDA approach as unsuitable for short-form text, proposing a hybrid model, which utilises mechanisms typical for the field of information retrieval. Another limitation, recognised by Dohaiha et al., (2018) is that by using LDA, topics require manual evaluation and are unlabelled, which offers potential for further automation. Considering the above listed limitations of the LDA method on short-form text, Chen et al. (2019) have compared its performance with the Non-negative matrix factorization (NMF) model, demonstrating that the latter is likely to perform better than LDA under the same configurations in topic mining for short texts.

LDA hybrid (sLDA) has also been developed for geo-aware topic models, suitable for offline analysis (Lozano et al., 2017). Such a tool has potential applications in consumer behaviour analytics. Another such relatively unexplored, but potentially impactful for understanding of cross-national consumer behaviour model is multilingual topic modelling. In this field, both LDA-based (BiLDA, bilingual-LDA) (Vulić et al., 2015), and hybrid (Lo et al., 2017) approaches have been proposed, the latter being based on unsupervised learning using a K-means clustering algorithm.

2.3.3 Other approaches: Applications and Limitations

Except for LDA, there are numerous other developments in the field of topic discovery. However, considering the lack of academic attention they have received, they appear to have critical limitations that remain unaddressed, as will be illustrated below. For example, Chen, Zhang et al. (2017) propose a hierarchical approach for topic detection where words are treated as binary variables and allowed to appear in only one branch of hierarchy. Although efficient when compared to LDA, it can be argued that this approach is unsuitable for application on short-form text, extracted from social media, considering the language ambiguity, which characterises this data form. Similarly, a Gaussian Mixture Model can be used for topic modelling of news articles (Jiang et al., 2018). This model aims to represent text as a probability distribution as means to discover topics (Jiang et al., 2018). Although it outperforms LDA, considering the lack of structure and data sparsity of short-form texts, it can be argued such a model will perform less coherently in topic discovery. Another model based on Formal Concept Analysis (FCA) was proposed for topic modelling of data from Twitter (Cigarrán et al., 2016). This approach shows facilitation of new topic detection based on information coming from previous topics, yet fails to generalise well, meaning that it is unreliable and sensitive to topics, which it has not been trained on.

Other models, such as Chen, Guo et al.'s (2017) TG-MDP (topic-graph-Markov-decision-process), consider semantic characteristics of textual data, as well as automatically select optimal topics set with low time complexity. Such an approach is suited for offline mode topic detection alone, yet shows promising results when compared to benchmark algorithms, based on LDA, which are considered superior to others in the field, such as GAC (see Yang et al., 1998), LDA-GS (see Asuncion et al., 2009) and KG (Sayyadi and Raschid, 2013). Finally, Dang et al. (2016) propose a dynamic Bayesian networks approach, which aims to detect emerging topics in microblogging communities. This field has more recently been furthered by Abulaish et al. (2018), who propose a five stage, topic evolution word embedding-based modelling approach, which analysis user-centric tweets to observe their topical evolution over a period of time. Although no research is found that builds upon this knowledge, these studies present possibilities to track the evolutionary behaviour of different user groups overtime, which can be useful for marketing strategists in determining the evolutionary direction of user interests.

To recap, although there are many approaches to topic modelling, LDA has evolved in being the most commonly used. Nonetheless, considering the model's limitations, a plethora of hybrid approaches have been subsequently developed to improve topic accuracy and relevancy, with methodologies being tested that challenge the model's probabilistic nature (e.g. hierarchical). Other non-LDA approaches have also been developed, however some limitations of their

application to short-form text are identified. Section 2.6 will further the discussion with an overview of methodologies developed specifically for short-form text, but first, an overview of the research of Sentiment analysis is provided in Section 2.4. and Challenges of both areas in Section 2.5.

2.4 **Sentiment Analysis**

2.4.1 Concept Overview

Sentiment analysis is a discipline that aims to extract qualitative characteristics from user's text data, such as sentiment, opinions, thoughts and behavioural intentions using NLP methods (Heimann and Danneman, 2014), with developments in the latter being highly relevant for the purpose of this research project's task. Social media texts are particularly useful for such type of research as they are used to express a standpoint, which is traditionally filled with subjective text (Zhang et al., 2018). Traditional studies on sentiment analysis have the aim to detect polarity in a given text, namely classifying it as positive, negative or neutral (Ravi & Ravi, 2015; Chen et al., 2018; Fan and Gordon, 2014). This categorisation need is considered one of the key limitations to traditional sentiment analysis, as subjectivity and objectivity are not addressed (Chaturvedi et al., 2018). More advanced methods attempt recognising multiple differentiated affective manifestations in text, which indicate emotions and opinions through analysis of the language used for self-expression (Sintsova and Pu, 2016; Chen et al., 2018). Additionally, such methods often aim to simultaneously detect and extract topic models, thus deep learning approaches such as convolutional neural networks (CNN) are often used (Wang et al., 2015; Wang et al., 2016). CNNs are also used in sentiment analysis of short-form texts (see Dos Santos and Gatti, 2014; Kale et al., 2018; Tang et al., 2015). The effectiveness of the sentiment extraction in short-form text relies on the application of more advanced methodologies (Dos Santos and Gatti, 2014). Social media data in particular requires comparatively more complex methods in information retrieval as well due to the creative language, use of slang and abbreviations (Baziotis et al., 2017).

Models used can vary between supervised (see Li, Guo et al., 2018; Bravo-Marquez et al., 2014), semi-supervised and unsupervised, with the former being most challenging to obtain and cost-inefficient for research (da Silva et al., 2016). Semi-supervised approaches utilise a small number of labelled samples as training data as means of improving classification accuracy, with an example being the model published by da Silva et al. (2016), where Twitter data is classified using SVM as an approach with resulting promising performance.

Sentiment analysis can be performed at a document level, sentence level and aspect (word) level (Diamantini et al., 2019). Short form texts, such as content from social media are best analysed with sentiment analysis at a sentence level as they usually consist of a single or few sentences

(Diamantini et al., 2019). However, models have also been proposed that analyse individual words under the assumption that words in the same sentence share the same emotion. Such an approach is Tang et al.'s (2019) hidden Topic-Emotion Transition model, which models topics and emotions in successive sentences as a Markov chain. This approach enables simultaneous detection of document-level and sentence-level emotion.

Multimodal sentiment analysis has grown as a field in recent years, with models proposed in the area taking advantage of recent developments in weakly supervised deep learning approaches (see Majumder et al., 2018; Chen et al., 2018). Simultaneously, multimodal event topic modelling has also emerged, which has been demonstrated as promising for the area of predictive analysis of consumer behaviour and sociology (Qian et al., 2015). Collectively topic modelling and sentiment analysis in a multimodal context are recognised as means of improving human-agent interactions, with an example being automatic speech recognition (Echeverry-Correa et al., 2015; Clavel and Callejas, 2015).

2.4.2 Primary Methods: Applications and Limitations

Sentiment analysis has initially been performed using pre-developed, manually built sentiment lexicons, such as Subjectivity Wordlist (Banea et al., 2008), WordNet-Affect (Strapparava and Valitutti, 2004), SentiWordNet (Baccianella et al., 2010; Appel et al., 2016), SenticNet, AFINN, Sentiwords, SO-CAL, Opinion lexicon, and WordStat, each having a various scale of rating and various word count (see Li, Guo et al., 2018). Such lexicons have been used as foundations for model development, with examples being the Polarity Classification Algorithm (PCA), which classifies tweet sentiment, the Enhanced Emoticon Classifier (EEC), Improved Polarity Classifier (IPC), and SentiWordNet Classifier (SWNC), amongst which superior performance demonstrates the PCA (Khan et al., 2014). These approaches although useful in distinguishing subjective or objective speech and categorising sentiment as positive, negative or neutral, enable researchers to extract sentiment primarily from the perspective of the writer as opposed to the reader (Rao et al., 2016).

Except lexicon-based approaches, sentiment analysis can be performed using a machine learning approach, which uses statistical models trained on human annotated datasets, thus utilising semi-supervised learning (Diamantini et al., 2019). Each perspective offers its own limitations and opts for compromising either accuracy of generalisability of the analysis. Almeida et al.'s (2018) study approaches the problem of multi-label sentiment classification from the perspective of the reader, applying a model to a news dataset. Their study demonstrates the superiority of ensemble classifiers when compared to other methods, providing a foundation for experimentation with such models on short-form text data. Table 2.4-1 (below) shows a comparison of the primary methods

used for determining semantic subjectivity in texts, alongside the advantages and disadvantages for each approach. For sentiment analysis of tweets, following a comparative analysis of six shallow machine learning approaches, Ahuja et al. (2019) conclude that TF-IDF perform better as compared to N-Grams in terms of feature extraction. Holistically, the combination of TF-IDF with logistic regression is considered most efficient amongst the studies sample of Ahuja et al.'s (2019) paper.

Table 2.4-1 Comparison of subjectivity detection semantic methods (Overview) (adapted from Chaturvedi et al., 2018)

Method	***Model***	***Advantages***	***Disadvantages***
Conditional Random Fields (Mao and Lebanon, 2007)	Sequence tagging, such as part-of-speech tagging and shallow parsing	Captures word order and grammar well (through n-grams)	High feature dimensionality
Semi-Supervised Learning (Pang and Lee, 2004)	Small number of labelled words of a known polarity are used for training and classification is done on highly similar samples small samples	Easy and time-efficient determination of polarity	Lack of in-depth understanding of subjectivity and objectivity in sentences
Deep Learning (Chatuverdi et al., 2016)	Input sequence processed by numerous layers, trained using backpropagation	Meta-level feature works well with large vocabularies, performs better than n-gram models	Does not perform well on short-form text and social media data due to noise in training data
Multiple Kernel Learning (Bucak et al., 2013)	Features organised intro groups, with each group having its own kernel function	Multimodal sentiment analysis	Slow computation

2.5 **Challenges of Short-form text Topic Modelling and Sentiment Analysis**

There exists no common definition on what short-form text is in academic literature, with scholars working with datasets, containing textual information from varying length with some examples of such data being user product and service reviews, textual data from Twitter (otherwise referred to as user Tweets), comments in public forums (e.g. Reddit), user posts from Facebook, comments on videos, and so on. Additionally, such texts can be instant messages, short message exchanges, forum comments and news headlines (Rao et al., 2016).

Short text is challenging for the tasks of topic detection and sentiment extraction as it lacks contextual information, which leads to a problem of data sparsity (Chen et al., 2011; Rao et al.,

2016; Ittoo et al., 2016). As a result, general models such as bag-of-words become unsuitable for semantic analysis of short texts as they ignore order and semantic relationships between words (Sriram et al., 2010; Tang et al., 2019). Nonetheless, a review of text analysis studies in financial markets demonstrates that the bag-of-words approach is used in the majority of the reviewed sample as means of feature selection (Nassirtoussi et al., 2014), which affirms its popularity in the academic community.

Currently the topic model quality depends manipulation and refinement, which is often manual and requires time-consuming fine-tuning of model parameters (Lee et al., 2017). One of the most considerable challenges in topic modelling is the issue of configuration. Prior to running a topic modelling algorithm, data pre-processing should occur, a step from which involves removing stop words and topic general words (TGWs), the latter traditionally done manually and considered a challenge in the research area. TGWs are problematic as they can alter the results of topic modelling as they are more probabilistic to occur in the corpus, thus more likely to be paired with other words, reducing the validity of word pair topics identified (Xu et al., 2017). Models have been developed to automate this task, which as a result is considered a means to improving the effectiveness of the topic modelling algorithm (Xu et al., 2017). Li, Zhang et al. (2018) propose the entropy weighting (EW) scheme, which is based on conditional entropy measured by word co-occurrences, combined with existing term weighting schemes, which can automatically reward informative words and as a result assign meaningless words lower weights, improving topic modelling performance. Lee et al. (2017) discuss how human interaction with topic models can also be considered another research challenge, proposing, following two individual experiments with non-expert users, that human-in-the-loop topic modelling is developed as a form of mixed-initiative interaction, where the system and the user work collaboratively with the goal of topic model optimisation.

Sentiment analysis on the other hand is primarily challenged by large datasets (Fernández-Gavilanes et al., 2016), which are often unstructured (unlike classical data mining corpuses) (Oza and Naik, 2016; Curiskis et al., 2019) and not annotated (Curiskis et al., 2019), thus are more difficult and time-consuming to pre-process for surface level machine learning. Choosing efficiently the pre-processing technique is considered a research priority, with studies being devoted to the topic, showing through comparative analysis means to improve the effectiveness of sentiment classification when using Tweets as data (Symeonidis et al., 2018). Twitter and other social media also present a challenge of irrelevant data collected as part of the dataset, which impacts performance of the model (Hajjem and Latiri, 2017). Liang et al. (2018) further argue sentiment analysis using topic-level and word-level models, which analyse short-form text are vulnerable to overfitting as a result of data sparsity. Additionally, microblogging involves using

flexible language, including abbreviations and slang as opposed to structured sentences, which is considered more challenging than traditional text for algorithmic analysis (Zhang et al., 2018; Ittoo et al., 2016; Khan et al., 2014; Appel et al., 2016). Part of the challenges in language interpretation are also the use of sarcasm, imagery, metaphors, similes, humour and figurative language, which relies on previous knowledge and/or context (Khan et al., 2014; Appel et al., 2016) as they impact sentiment classification accuracy. The lack of gold standards and annotated data in the fields of topic modelling and sentiment analysis result in reduction of the academic rigour of many studies due to subjectivity and ambiguity (Ittoo et al., 2016). Annotation in itself is time-consuming and complex (Ittoo et al., 2016), which is why the majority of studies deploy unsupervised learning algorithms.

2.6 Topic Modelling and Sentiment Analysis of Short-Form text

When performing social media sentiment classification tasks scholars approach the classification task from a semi-supervised perspective, equipping the model with a sentiment dictionary, which includes relational conjunction, emoticon, negative word, network word, basic sentiment and degree adverb dictionaries, which collectively enable apt decision-making (Zhang et al., 2018). Such an approach although time-consuming addresses the joint requirements of both topic modelling and sentiment analysis in short-form text, however, it can be criticised for overreliance on manual class definition and little automation. Arguably, such an approach would be hardly generalisable or scalable.

As a result, deep learning techniques have increased in popularity in the field, considering they offer automatic feature extraction and both richer representation capabilities and better performance, when compared to surface models (Araque et al., 2017). Yet, considering previous analysis on deep learning on short-form text (see Table 2.4-1), the importance of noise-reduction in training data is vital for performance optimisation. Some models have already been developed, which use convolutional neural networks (CNN) for short text modelling, showing comparative accuracy superiority to other models (Wang et al., 2016). Although the sentiment classification problem can be solved using surface learning models (e.g. SVM) (Bhadane et al., 2015), the superiority of deep learning approaches (e.g. deep neural networks) for sentiment classification task is shown to outperform models such as SVM or NB in comparative analysis (Sun et al., 2016). Nonetheless, it is to be noted that base learner architectures can be improved using ensemble methods (bagging, boosting and random subspace), as demonstrated by Wang et al. (2014).

Recurrent neural networks (RNNs) are also considered suitable as means to solve the challenges short form text poses for sentiment analysis and topic detention as such networks have memory

capabilities, which can be utilised to process the input in a sequential manner as opposed to a bag-of-words, as mentioned above (Abid et al., 2019). Abid et al. (2019) and Rosa et al. (2018) propose an architecture that utilises the advantages of both CNN and RNN (Recurrent Neural Network) through layers of a deep learning network in combination with other functions, demonstrating reliable classification accuracy and improved structure in terms of less required layers and processing. A review of deep learning approaches used for sentiment analysis, extracted from product reviews shows recurrent neural networks to be most common amongst research approaches, followed by CNN and recursive neural networks (RecNN), alongside a plethora of hybrid approaches, which include variants of Long-Short Term Memory (LSTM) (see Zhang et al., 2019), Gated Recurrent Unit (GRU), pre-trained and fine-tuned word embeddings, and incorporated linguistic factors in the form of part-of-speech and grammatical rules (Dohaiha et al., 2018). Such models are determined to still be in its relative infancy when compared to more traditional shallow machine learning approaches. Nonetheless, considering the popularity of RecNN models in recent years, comparative research experiments have been carried out testing means to improve their performance, including through ensemble techniques for deep learning models (Araque et al., 2017). Table 2.6-1 below summarises key advantages and limitations to each of the discussed methods.

Table 2.6-1 Comparison of Deep Learning methodologies (adapted from Dohaiha et al., 2018)

Method	***Advantages***	***Disadvantages***
CNN	Ability to extract meaningful local patterns (n-grams) Non-linear dynamics Time-efficinet Computation	Extensive Preprocessing requirements Hidden layers limited in terns of size
RNN	Distributed hidden states can store past computations Does not require a large dataset	Requre fewer parameters Potential for false prediction Fails to capture long-term dependancies
RecNN	Simple architecture Learns tree-like structures Can construct representations for new words	Requires extensive parameters Prone to inacuraccies Lack of research

User reviews have been used in several papers as an example of classification on the basis of text aspects and sentiment, with a model suitable to perform this function is LSTM with aspect-embedding and text autoencoding (Fu et al., 2019). Comparatively, Li et al. (2019) propose a joint sentiment-topic (JST) model for analysing user reviews, which collectively achieves the goal of analysing sentiment and identifying topics, which are most critical for the analysed text, specifically demonstrating how such a model can be used in a sales improvement context. Ali et al. (2019) utilise a variety of data sources, including long and short-form texts collectively and by using a pre-trained word embeddings model, achieve better sentiment classification performance on topic models.

Measuring emotions in readers is less commonly addressed in academic literature, and arguably more challenging. Through an architecture of unsupervised learning or topic-level maximum entropy (TME), Rao et al. (2016) measure social emotion classification of short-form text. Such an approach is demonstrated as useful for the purposes of marketing intelligence. Comparatively, Liang et al. (2018) propose a model that that performs short text classification from a reader's perspective by introducing a topic-emotion layer. Their model, however, fails to outperform other models in the conducted experiment, specifically lagging on classification accuracy.

Chen et al.'s (2018) study demonstrates a model that analyses sentiment from short-from exchanges (i.e. chat messages), also taking advantage from the use of emojis as means to enable more accurate emotion recognition in informal messages. The study also demonstrates a practical application of such a system by carrying out experiments with users testing a prototype system that performs the analysis in real-time. Alongside emojis, the user's personality characteristics can be extracted as means of more efficient sentiment classification (Huang et al., 2017), as well as the sequence of the sentences (Qiu et al., 2018). Hashtags in combination with emojis are also considered very efficient for classification (Howells and Ertugan, 2017). Additional factors, such as connections in social media networks are also considered as influencing factors regarding the analysis of sentiment of microblogging publications (Xiaomei et al., 2018), however community detection is a novel stream of literature with few papers published on the topic.

Other approaches were also identified, which can be commonly grouped into two categories, LDA-based and non-LDA using, summarised in consideration of their methodology and used data (see Table 2.6-2, below). The combination of approaches and the hybrid nature of most models, illustrated in the table address the limitations of the methods, which were discussed in previous sections.

Table 2.6-2 Overview of Extracted Models, Methods and Applications for Simultaneous Topic Modelling and Sentiment Analysis of Short-form text

LDA-based Approaches		
Model	***Method***	***Used data***
MJST *(Huang et al., 2017)* *multimodal joint sentiment topic model*	LDA, data from emoticons, publishers' personality	Tweets
NHDP (Fu et al., 2015) *non-parametric hierarchical Dirichlet process*	hierarchical Dirichlet process with a semantic layer	Social media posts
IG and LDA-IG (Zhang et al., 2016) *IdeaGraph and Latent Dirichlet Allocation*	graph analytics	News documents, Tweets
Ontology and LDA *(Ali et al., 2019)* *Latent Dirichlet Allocation-based topic modelling*	Word embeddings	Tweets, Online reviews, news, Facebook posts
UTSJ *(Dong et al., 2018)* *unsupervised topic-sentiment joint probabilistic*	LDA with added sentiment level, Gibbs sampling	User reviews
Non-LDA Approaches		
WS-TSWE *(Fu et al., 2018)* *Weakly supervised topic sentiment joint model with word embeddings*	word embeddings HowNet lexicon Gibbs sampling algorithms	Online reviews
WSTM (Xiong et al., 2018) *Word-pair Sentiment Topic Model*	Gibbs sampling	Product reviews
Union model *(Ren et al., 2016)*	SVM bag-of-words, sentiment lexocons, PMI unigram lexicons, PMI bigram	Twitter

	lexicons, negation detection, elongated words	
Bayesian model (Farhadloo et al., 2016)	Bayesian model with Low dimensionality	TripAdvisor user reviews

2.7 **Identification of a Research Gap**

Recognising the limitations of published research is considered vital for providing readers with an accurate representation of the academic knowledge on the topic (Booth et al., 2016). The following section will present a critical assessment of analysed evidence, with the aim of recognising any collective gaps of knowledge, which the current research can address.

As demonstrated by the previous section, there is a considerable number of significant studies that approach both topic modelling and sentiment analysis of short-form text (see Rao et al., 2016; Diamantini et al., 2019; Li et al., 2019; Ali et al., 2019; Huang et al., 2017; Zhang et al., 2016; Dong et al., 2018; Fu et al., 2018; Xiong et al., 2018; Ren et al., 2016; Farhadloo et al., 2016), which can be utilised as foundational knowledge in the process of system development. Lee et al.'s (2017) study is particularly relevant in that it demonstrates the necessity of user involvement in the process of model refinement and optimisation to improve accuracy and optimise output. Another highly influential piece is the study of Ibrahim and Wang (2019), where LDA analysis is used for topic modelling and subsequent sentiment analysis of Tweets, with the aim of evaluating retail service efficiency. The study demonstrates how business analytics through intelligent methods can be utilised for strategic improvements. Nonetheless, it is recognised that the models proposed as part of the paper can be improved considering the human involvement in topic identification and labelling (Ibrahim and Wang, 2019). Involvement of human agents is not desired as most industrial applications are moving towards full automation.

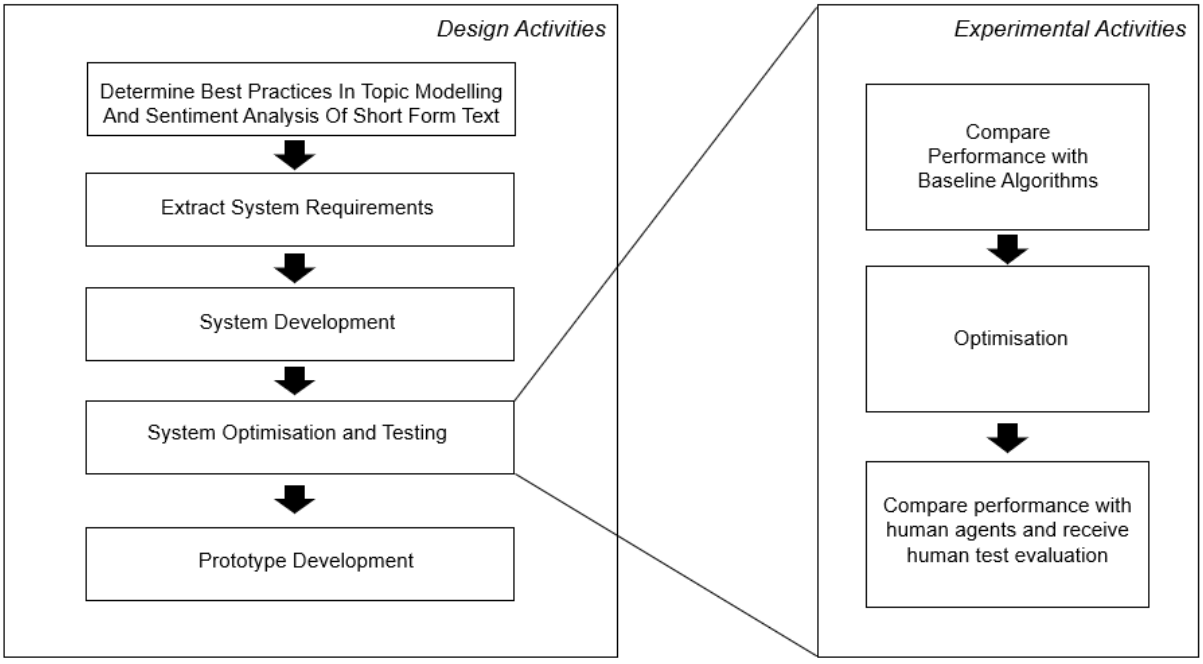
In general, very few studies form the reviewed sample have recognised the role of topic extraction and sentiment analysis as means of extracting user preferences, and subsequently optimising marketing strategy (see Rao et al., 2016; Li et al., 2019; Wang et al., 2018; Howells and Ertugan, 2017; Ravi and Ravi, 2015; Dang et al., 2016; El-Diraby et al., 2019; Farhadloo et al., 2016). Although the possibility of market prediction, using combined sentiment analysis and topic models has been analysed in the context of various industry settings, such as financial markets (Nassirtoussi et al., 2014), political events (Lozano et al., 2017), improvement of recommendation algorithms (Wang et al., 2018; Rosa et al., 2018; Xiao et al., 2019), retail (Ibrahim and Wang, 2019), location-based sociological analysis (El-Diraby et al., 2019), airline service quality (Korfiatis et al., 2019) and social trends and viral topics (Li, Wu et al., 2018). However, it is recognised that

no studies have approached classifying topic models and sentiment of users as stages of their customer journey. Considering the potential marketing and business applications of this solution, it is proposed that the current research will address this knowledge gap. The following section will demonstrate the conceptual model that will be applied further.

2.8 **Conceptual Model**

The conceptual framework provides the methodological bases for development of the proposed experiment and system and analysis of findings. A theoretical model, i.e. a detailed analysis of what the proposed system should include will be provided in the following chapter. Figure 2.8-1 (below) demonstrates the conceptual model of this research project and will be used as reference for the system development process and associated experimental activities. Next will be presented the Methodology chapter, which will demonstrate the protocols and procedures for the system development and associated experiments.

Figure 2.8-1 Conceptual model of proposed research (methodological)



2.9 **Conclusion**

The Literature Review Chapter had the aim of reviewing recent publications in the areas of topic modelling and sentiment analysis of short-from text in a systematic manner, as well as discussing them in a critical manner, identifying and demonstrating the knowledge gaps that exist. As a result, a rigorous search strategy was developed, addressing these requirements. In total, 100 studies were reviewed, 78 of which identified following the database search and the rest through hand searching and reference list searching. The literature review demonstrated the suitability of

exploring topic modelling and sentiment analysis of short-form user or consumer generated text as means of customer journey stage classification.

3. CHAPTER III: METHODOLOGY

3.1 **Introduction**

3.1.1 Research Questions and Hypotheses

In Section 1.5 (Chapter I: Introduction), the research questions were introduced. Prior to examining the applied methodology in-depth, a short discussion of the hypotheses that relate to each question will be provided.

Question 1 asks: '*Can a library-generated sentiment classifier replace a manual sentiment classification process efficiently?*'. To address this question, the TextBlob library tool will be applied for sentiment analysis of the dataset. Its performance will be compared with the sentiment evaluation made by study participants, the results of whom will stand for manual sentiment classification. An alternative (lexicon-based) sentiment classification methodology will also be proposed. Although research directly addressing this research is not found, it can be argued based on other study results, that the current hypothesis supports the application and training of more advanced classification algorithms, e.g. ensemble learners as opposed to using the library-based approach alone (Yan et al., 2017; Srinivasa-Desikan, 2018). This can be considered a null hypothesis, which will be opposed.

Question 2 asks: *'Which topic modelling technique can be considered most efficient for handling of short-form, user-generated social media textual data, with experimentation comparatively evaluating the performance of LSA and LDA for topic coherence and similarity with topics generated by humans on a small sample of the data?'.* Previous research suggests that the LSA model is superior to LDA, when analysing movie reviews (Bergamaschi and Po 2014), which are longer of form, therefore that this can be taken as a null hypothesis. An opposing hypothesis is that LDA demonstrates superior performance on short-form text, affirming why it is so often chosen as a model for topic modelling academic research (see Section 2.3.2, Chapter II).

The final question is: *'Which classification technique can be efficiently applied to a web-extracted dataset with user-generated text to categorise the data entries into five categories that correspond with the user journey?'.* The experiment involves comparative evaluation of supervised and potentially unsupervised model, depending on the performance of the former on web-extracted data. Logistic regression, Support Vector Machines (SVM) and Naïve Bayes are commonly mentioned in academic research for topic-sentiment classification (see Gupta et al., 2017; dos Santos and Ladeira, 2014), hence why they will be developed for supervised models. Considering the growth of using deep learning methods for working with short-form text, demonstrated in Chapter II, deep learning architectures will also be proposed. However, as the data that this study

will be working with is unlabelled, the application of an unsupervised (clustering) and semi-supervised (lexicon-based) approach will be evaluated in contrast with manual classification.

3.1.2 Deliverables

The recent developments in the online availability of scientific knowledge and data, driven by the move from print to online publication of academic research, greatly support the formalisation of science (Soldatova and King, 2006). Although the traditional presentation of findings in the form of written natural language is still necessary, online publications allow authors to support the validity of their arguments through presenting the formal experiment data, publishing all data and associated metadata of a scientific experiment for posterity, allowing experiment repeatability and comparative analysis (Soldatova and King, 2006). Therefore, the intended deliverables include not only the written thesis, but also system development and optimisation code, analytics data from performance testing and participant testing, as well as a demo of the final model (post-evaluation) in an screen-capture format, the latter of which will be presented in Chapter V.

3.1.3 Chapter Structure

This chapter aims to clarify, justify and rationalise all research design decisions, the ultimate purpose of which is to answer the research questions as clearly and efficiently as possible (Bloomberg and Volpe, 2018). First a research overview will be provided, in Section 3.2 discussing the research philosophy, paradigm and strategy, with techniques and procedures explained in detail in Section 3.3. Limitations, ethical considerations and an appraisal of the academic rigour of the research will be discussed as well towards the end of the chapter.

3.2 **Research Overview**

3.2.1 Research Philosophy

The research philosophy demonstrates how knowledge is deemed as such and how the research process is conceptualised. The rationale for illustrating the research philosophy is the impact it has on both the way research is designed, as well as how results are interpreted. There are broadly two approaches to knowledge: objective and subjective, where the former is scientific and results-oriented, and the latter is exploratory and reason-searching (Davidson, 1996). Objective research is thus separated from society and does not concern itself with its diverse reality, but instead only serves to provide an understanding of raw data as opposed to subjective research where the cause and implications of a phenomenon are explored in-depth (Cooper and Schindler, 2014). Objective research is transparent, highly accurate and often considered the more scientific of the two philosophies (Sarantakos, 2012), yet it is recognised that each research philosophy has its merits depending on the examined context. Objectivity is also considered to provide a

modest, focused understanding of a subject matter, as opposed to subjectivity, which aims to answer more broad questions. The philosophy of the current research is objective as it aims to answer the proposed questions in a quantitative, data-informed manner through a strategy of positivism, where a systematic observation of facts is made and logical reasoning is applied to interpret the data and findings, form and test hypothesis and report results (Quinlan et al., 2019). As a result, the impact of internalised beliefs and the cultural background of the researcher have no impact on the research design, execution and result interpretation – something, which is otherwise (i.e. in subjective research) commonly mentioned as a limitation (Chiu et al., 2010; Flick, 2014).

3.2.2 Research Paradigm

The research paradigm is the means of perception, otherwise a representation of beliefs and values in disciplinary research (Schwandt, 2001; Saunders et al., 2016). As such it guides the methodology of solving the problem at hand. There are several components of a research paradigm, each offering to the researcher various options as to how the research should be approached (Johnson and Onwuegbuzie, 2004).

Part of the research paradigm is the ontology and epistemology. The ontology explains the researcher's philosophical assumptions regarding reality in a social context (Goodson and Phillimore, 2004), whereas the epistemology refers to how knowledge is established as such (Patton, 2002). Three common world views (i.e. ontologies) are accepted in scholarly research: constructivism, objectivism and pragmatism (Goodson and Phillimore, 2004). The current research is conceptualised, planned, executed and interpreted under the philosophical assumption of objectivism, which separates the subject from the object, treating research as means of uncovering a universal, objective truth (Bernstein, 2011). The impact this has on the current research is that user-generated texts will be considered as data points, with the meaning and impact of user stories remaining unexplored and unaddressed. The epistemology centres around the understanding of knowledge, and how beliefs are justified and rationalised (Norris, 2005), as well as what knowledge is deemed by the researcher as sufficient to answer the research questions (Saunders et al., 2016). Knowledge is established as such following statistical and quantitative validation, which is considered an aim at each stage of system development, as will be demonstrated further. Nevertheless, considering the scarcity of quantitative model evaluation techniques when working with unlabelled data, it is anticipated that model evaluation will require a degree of non-scientific result interpretation, i.e. qualitative analysis.

The last part of the paradigm is the axiology, which generally explains the role the researcher plays in influencing the written piece, specifically the researcher's awareness of how their values

and opinion might impact the reporting of results. It is recommended that the axiology is discussed as it assists in improving the transparency between the research author and the reader (Saunders et al., 2016). The researcher's background in business studies and marketing influences the frame of research being more focused on demonstrating a real-world application to the proposed system, with it solving challenges that are present in the current marketing analytics process for companies such as MyCustomerLens (Appendix A), as well as in other industries, such as learning analytics, tourism, retail, which will be expanded on in Chapter V. It is considered that the demonstration of academic merit is within the complete system prototype development, which includes experimentation procedures, as well as the demonstration of implications for businesses following the introduction of the proposed short text user-generated data analytics system.

Collectively, the decisions, related to the research paradigm indicate that the reasoning applied in the research process is inductive as opposed to deductive (Saunders et al., 2016; Feeney and Heit, 2007), with knowledge acquired through the stages of conceptualisation, modelling and analysis (Dodig-Crnkovic, 2002). The next section addresses the resulting research strategy.

3.2.3 Research Strategy

The adopted research strategy is an experiment as experiment designs are considered to be more rigid and scientific from a structural perspective, which enables replicability and greater validity of research (Saunders et al., 2016). As a result of the conducted experiment, data will be generated that illustrates the superiority of a combination of models, which can be used for comparative evaluation with manual task completion and analysed with reference to other research data (Tichy, 1998).

Although a mono-method quantitative methodology is the natural extension of the research philosophy and paradigms explained above, considering the lack of labelled testing data, performance will be evaluated through a mixed methodology through cross-comparison with the performance of study participants. Qualitative interpretation of findings is necessary for the interpretation of topic models as a first instance of verification of topic coherence on unlabelled datasets, as affirmed by scholars in the field (Chuang et al., 2012; Wallach et al., 2009). The study's time horizon is cross-sectional (Saunders et al., 2016), which is defined with the data, analysis and reporting being done at a single point in time. A resulting limitation is the lack of adaptability of findings to change, however, this risk is inevitable in fast-paced and dynamic research fields such as machine learning and NLP. In the following section, the techniques and procedures will be discussed in detail. To summarise the research ontology, based on Saunders et al.'s (2016) research onion is presented as Table 3.2-1, below.

Table 3.2-1 Summary of Research Ontology (adapted from Saunders et al., 2016).

Ontology Layer	*Choice*
Philosophy	Positivism
Approach	Deductive
Methodological Choice	Experiment
Strategy	Mixed Methodology
Time Horizon	Cross-Sectional

3.3 **Techniques and Procedures**

3.3.1 System Requirements

The required product functions, alongside their descriptions, functional and non-functional system requirements are summarised in the Table 3.3-1, below. This method of requirements reporting is consistent with IEEE Computer Society's (1998, reaffirmed in 2009) guidelines for Software Requirements Specifications, with the rationale and aim of the system being affirmed previously in Chapter I, and the unique specifications of various components being detailed further in this chapter. This requirements catalogue can be used as guidance for the final system, with the current research providing a prototype system, based on the same requirements as a result of the project's scope. Some notable out-of-scope activities for the current project are: (1) data labelling and (2) dynamic (real-time) access to social media, the former of which will affect the text classification techniques used.

Table 3.3-1 System Requirements Catalogue Brief (adapted from IEEE Computer Society, 1999; 2009)

Function	*Description*	*Functional Requirements*	*Non-functional requirements*
User-generated Social media Text Extraction	Extracting public data from a social media platform	Enables dynamic data mining with access to public social media API	Usability Reliability Performance Coherence
Text pre-processing	Pre-processing data to allow application of machine learning models	Cleans data from noise, inaccuracies, stop words and other frequent and rare words	

		Applies advanced feature extraction procedures that can be used by machine learners in subsequent stages
Sentiment Analysis	Identify the sentiment polarity of a given text	Classifies sentiment polarity of texts
Topic modelling	Extract key topics that are coherent and usable for identifying key areas of concern for the business	Requires an efficient, fast system that extracts coherent topic models
Text Classification	Classify text as one of five categories of user behaviour	Requires system for manual/automatic data labelling for model training Requires a well-trained classifier to categorise texts with expert knowledge in relevant classes

3.3.2 System Development

3.3.2.1 Data mining and associated procedures

Littman (2017) identifies four primary ways of acquiring Twitter data: retrieval from the public API (Application programming interface), finding an existing Twitter dataset, purchasing from Twitter and access or purchase from a Twitter service provider. Due to the lack of sponsorship associated with the project, no cost-associated activities are intended. Therefore, data was obtained through Twitter and Facebook's public APIs through using an online data scraping service – *Netlytic* (Gruzd, 2016).

Prior to explaining the procedure for data extraction, a rationale for choosing the social media platforms will be provided. Twitter is one of the most popular social media platforms, with research showing that users post more than 500 million tweets daily on average (Crannell et al., 2016; Öztürk and Ayvaz, 2018). The platform's user base also accounts to and exceeds 22% of the internet users of the world, which offers an opportunity for instant, real-time market insight (Kayser and Bierwisch, 2016). Together, Facebook and Twitter are considered the most 'crowded' social media platforms and are thus most commonly used in social media analytics and NLP research (Salloum et al., 2017). With the growth of popularity of these platforms, their use has transitioned from solely a platform for sharing life updates with friends and family to a tool for direct

communication with companies, word-of-mouth and network marketing (Rybalko and Seltzer, 2010; Einwiller and Steilen, 2015), which is especially relevant for the current research. Such data is an opportunity for NLP as it represents a digital trace of B2C and C2C communication, which can be utilised to create an overview of the user, their relationship with the organisation, and their user journey, all based on their publications.

Netlytic (Gruzd, 2016) extracts data through social media website's public APIs, specifically Twitter's REST API v1.1 and Facebook's Graph API v2.7 (Gruzd et al., 2016; Gruzd, 2016). A limitation of this service is the limits that exist for data extraction, i.e. up to 1000 tweets per query for Twitter, and up to 2500 posts for Facebook, with the latter returning posts and replies for public Facebook groups, pages, events or profiles. With Facebook, however, a further limitation is applied of the 100 top level posts from a page, as well as up to 25 replies per post (Gruzd, 2016). Considering these limitations, the technique used to extract relevant information was keyword search, which is a popular data mining procedure for social media posts (Ampofo et al., 2015; Gruzd et al., 2016). This was complimented by searching for Facebook posts, where users have mentioned insurance company names or commented on insurance companies' corporate publications. The rationale for this being that a large number of users at various stages of their customer journey would use a corporate social media page (on Facebook) or profile (on Twitter) as a platform for direct interaction with the organisation (Rossmann and Stei, 2015; Einwiller and Steilen, 2015), with examples including asking specific questions about the company's prices (indicating the user is at the information search stage), the user's own policy or experience (indicating they are in a current relationship with the organisation) or advocating for or against using a company (indicating a post-evaluation stage). To see the complete procedure used for data extraction from Twitter and Facebook, including relevant keywords, search terms and companies, whose mention was specifically sought, please refer to Appendix Data Extraction Procedure: Limitations, Stats and Queries, TablesTable D-5.7-2 (for Facebook) and Table D-5.7-3 (for Twitter).

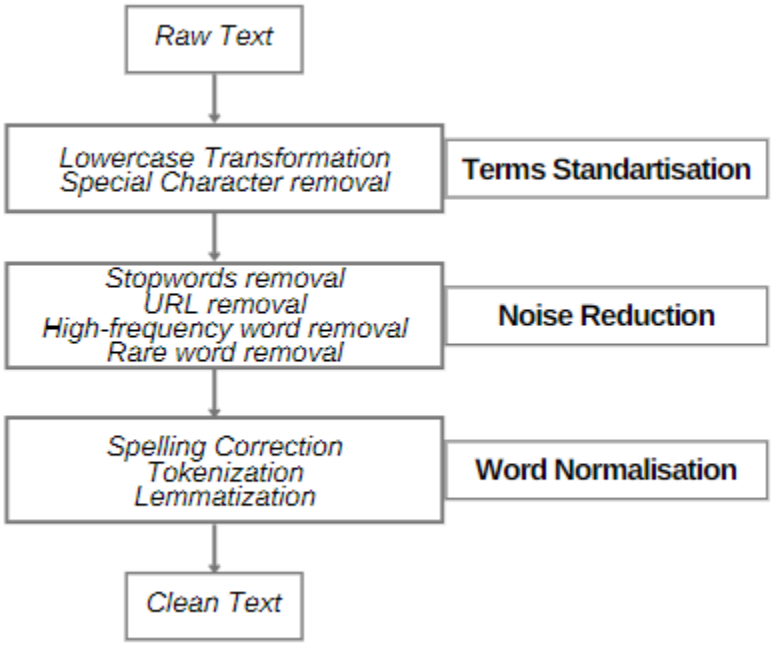
3.3.2.2Data pre-processing

As in the previous section it was explained that *Netlytic* extracts both the publications, and comments left on those publications, the data was first and foremost rid of corporate posts. In addition, although the service captures contextual information for the user, which can be used to demonstrate the user's influence and how their tweets can potentially impact the organisation or other consumers, such information has been stripped from the training dataset, with only the publications themselves being used in the final dataset of the current research. This also enables identity protection and anonymity of the users, whose posts have been included in the dataset.

With a similar goal, as well as to optimise the model's performance, usernames have also been removed from the corpus (Gupta et al., 2017). Retweets and duplicate posts have also been removed to reduce the noise in the data (Gupta et al., 2017). Following these procedures, the dataset size reduced from 44,207 individual text entries to 16,269 entries.

Although a great volume and variety of data is being generated daily, textual data extracted directly from social media websites is unsuitable for machine learning analysis unless it has been prepared for this purpose (Srividhya and Anitha, 2010). Although, it is worth noting that some studies that test the effect of pre-processing techniques on the performance of sentiment analysis models conclude that the pre-processing does not result in significant improvement of performance (dos Santos and Ladeira, 2014), generally it is considered that this step is important as it results in clarity of input data for the learning algorithm, consequently impacting the processing speed and the accuracy of output (Srividhya and Anitha, 2010). Exploration and preparation of data involves, but is not limited to writing functions for filtration of noise from data, setting up the development environment, scaling and encoding, guidance for which has been extracted from the academic texts of Géron (2017), Chollet (2018), Nielsen (2015), Goodfellow et al. (2016), Russell and Norvig (2016) and Berry and Linoff (2004). Such a process is commonly referred to as pre-processing and it broadly includes three main steps: term/object standardisation, noise reduction and word normalisation, each of which consists of various text analysis operations that must be performed (see Figure 3.3-1 Data Cleaning Pipeline, below). The methodology of each procedure will be described briefly in the following paragraphs.

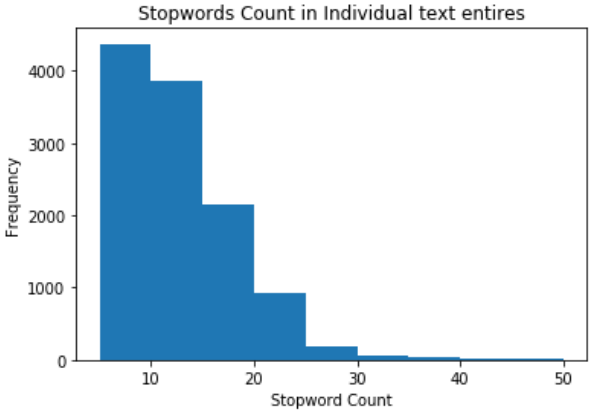
Figure 3.3-1 Data Cleaning Pipeline



The first step of pre-processing is transforming all user-generated text into lowercase. This is done to avoid the processing of the same words differently, e.g. ‘insurance’ and ‘Insurance’. To reduce the size of training data, punctuation and hyperlinks are also removed as they do not add any information that is valuable for the analysis (Sun et al., 2014). This process is referred to as terms standardisation (dos Santos and Ladeira, 2014).

Stopwords are commonly occurring functional words, which are frequently used but carry no information (e.g. pronouns, prepositions, conjunctions) (Srividhya and Anitha, 2010; Sun et al., 2014; dos Santos and Ladeira, 2014). In the English language, there are many such words, and filtering them allows for data handling and time efficiency (Hardeniya et al., 2016; Srinivasa-Desikan, 2018). To affirm the necessity of removing stopwords, the following Figure 3.3-2 is attached, which shows the prevalence of such words in individual dataset entries, with the majority of texts containing between 5-15 stop words each. The NLTK (Natural Language Toolkit) Python library has various packages that support text pre-processing, including a *stopwords* pre-defined library (purpose (Gupta et al., 2017; Sarkar, 2016), that has been utilised for the purpose.

Figure 3.3-2 Stopwords Count in Individual Text Entries



While stopword removal eliminates high frequency words in the general language, there might be high frequency (common) words in the dataset, which are contextual. Keeping such words in the dataset can lead to skewed results, with the assumption of this being that frequent words are not informative for category prediction (Srividhya and Anitha, 2010). Similarly, rare words in the dataset can be considered as outliers due to the association between them and other words being dominated by noise (Sarkar, 2016; Srinivasa-Desikan, 2018). Prior to removing the most common and rarest words in the dataset, they were looked in to, to affirm the rationale of this procedure (see Table 3.3-2 below). From the most common words, it is evident that the only word with significantly disproportionate frequency is ‘insurance’, the instances of which have been removed to avoid skewing of the models’ results. All words from the least common list have been removed.

Table 3.3-2 Most common and Least Common Word List

Most common words	*Frequency*	*Least Common Words*	*Frequency*
insurance	*13684*	problemâ	1
car	2481	predicted	1
policy	2481	httpstcocil0dfseb4	1
rt	1804	httpstcop5xzbgkvm3	1
life	1171	illustrate	1
Insured	1755	overs	1
Get	1683	ðÿeveryone	1
Health	1458	blogwhatever	1
new	1410	tytarmy3	1
Company	1350	takethatgravity	1
umbrella	1257	abha	1
Help	1245	philkhfc	1
Im	1217	fsahsa	1
Coverage	1125	ðÿµðÿ¹â	1
Amp	1112	sakz5th	1
Business	1021	prof_noface	1
Looking	993	ucl	1
One	970	driverside	1
would	962	testâ	1
Pay	933	crescenteagle	1

When working with user-generated social media text, research has demonstrated the importance of spelling correction, which helps reduce the multiple copies of the same word (Sun et al., 2014; Clark and Araki, 2011). To perform spelling correction of the data, the *Textblob* Python library has been used. A limitation of using this method is that it takes a long time to process the task, as well as such an approach not being as accurate when compared to manual correction. Nonetheless, it is a preferred method when working with high-volume datasets of user-generated social media data.

Tokenization is the process of dividing text into a sequence of words or sentences that carry meaning (Clark and Araki, 2011; Vijayarani and Janani, 2016). The *Textblob* library has been used for this again, which performs the function of transforming the text into a blob, and subsequently converting it into a series of words (Vijayarani and Janani, 2016). A limitation of *TextBlob* is that it cannot tokenize special characters (Vijayarani and Janani, 2016), however this was accounted for in previous steps.

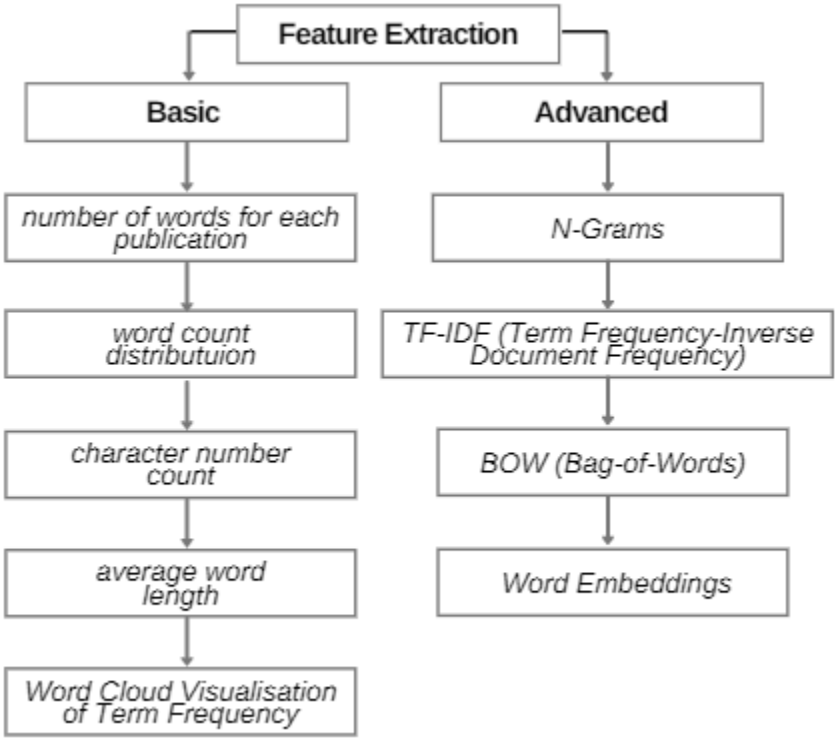
Lemmatization is a method that converts the word into its root word, which is considered a more effective pre-processing alternative to stemming, for example, which only removes the suffices of words and is more frequently used in academic research (see dos Santos and Ladeira, 2014; Sun et al., 2014). Lemmatization has a comparatively more advanced method as it uses a vocabulary

to perform a morphological analysis to obtain the root word and uses that word instead; however, studies show that it is traditionally used for linear text classification systems (Camacho-Collados and Pilehvar, 2017; Bird et al., 2009).

3.3.2.3Data exploration and Feature Extraction

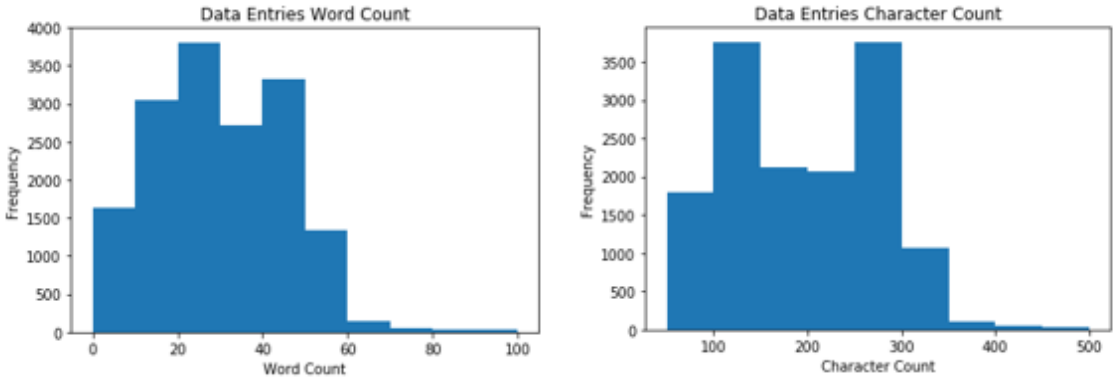
Although the operations, described in Section 3.3.2.2 are sufficient to prepare the text for a machine learning NLP model, additional steps have been taken for familiarisation with the dataset, which can later be used in the process of system fine-tuning and performance optimisation (Srinivasa-Desikan, 2018). These procedures can be summarised in two categories: basic and advanced feature extraction (see Figure 3.3-3, below).

Figure 3.3-3 Feature extraction pipeline



Basic feature extraction serves well for familiarising oneself with data, with notable aspects of the dataset following pre-processing demonstrating that the combination of data sources (i.e. Twitter and Facebook) resulted in a variety of texts in respect of the word count and character count in each text, as demonstrated by Figure 3.3-4, below.

Figure 3.3-4 Data entries individual count of words per entry (left) and characters per entry (right)



The most frequent words in the dataset have been plotted in the word cloud below (see Figure 3.3-5). It can be seen from the words plotted that there are various instances of words that express semantic characteristics, e.g. avoid, great, good, better, as well as such that can be potentially used for user journey class specification, e.g. looking, need.

Figure 3.3-5 Most Frequent words in the Dataset, represented in a Word Cloud Format

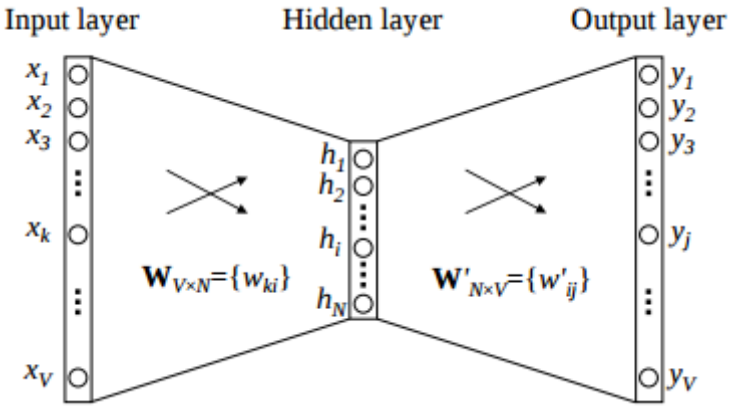
Considering the importance of advanced feature extraction procedures for training text classification algorithms, in the following paragraphs the rationale and potential application of the techniques shown previously in Figure 3.3-3 will be discussed.

The Term Frequency (TF) factor affects the importance of a term in the document and is frequently discussed in pair with Inverse Document Frequency (IDF) (Srividhya and Anitha, 2010; Bonzanini, 2016). TF is measured for each word and is a mathematical weight representation of the distribution of this word in the document, whereas IDF measures the frequency of each word in the text corpus (Srividhya and Anitha, 2010), or in the current case – collection of user-generated texts. Collectively, the TF-IDF ratio is very useful for text analysis that involves sentiment analysis as it penalises words that are frequently occurring such as 'don't' or 'can't', but instead gives high weights to words such as 'disappointed' since they carry contextual information, useful for determining the sentiment of the text (Thanaki, 2017). A *sklearn* function has been used to directly obtain TF-IDF vectors.

N-grams are combination of words, with N representing the number of words in the combination. N-grams, where N=1 are referred to as unigrams, with bigrams and trigrams representing combinations of 2 or 3 words, respectively (Gupta et al., 2017). Unigrams are less commonly used as they contain less information as opposed to bigrams or trigrams. The rationale for using n-grams is that they capture the means of expression and the language structure, which can as a result enhance the fit of the machine learning model (dos Santos and Ladeira, 2014). Consequently, longer n-grams contain greater contextual information that shorter ones, however considering the short nature of user-generated text on social media websites, n-grams longer than trigrams will not be representative of the majority. The *Textblob* Python library has been used to extract n-grams.

Word embeddings are vector representations of text. They are used with the aim of extracting patterns from the corpus, with the underlying idea being that words that are similar will have a minimum distance between their vectors. They are dense, relatively low-dimensional and learned from the data at hand (Chollet, 2018). The standard methodology of vectorisation is shown in Equation 3.3-1, below. *Word2vec* is commonly used in NLP, however a limitation that researchers often face is the lack of data to train the *word2vec* model on. As a result, pre-trained word vectors can be used for optimising model performance. Different vectors have been trained on wiki data in various dimensions, which are made publicly available through *GloVe* (see Pennington et al., 2014), which is often used in research to obtain global-level representation of words to summarization (Thanaki, 2017; Srinivasa-Desikan, 2018; Chollet, 2018). For the current research, the 100-dimensional version of the model has been used, which when trained on the social media dataset enables vector representations for specific words and phrases.

Equation 3.3-1 Vectorisation Methodology



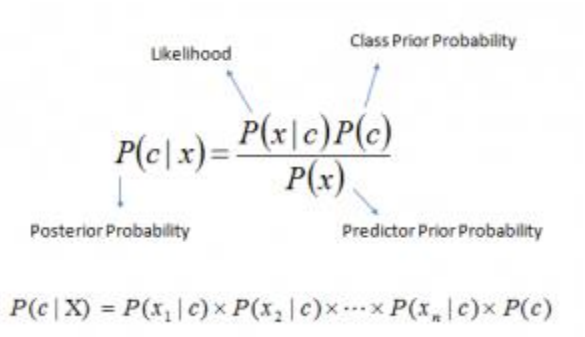
3.3.2.4 Sentiment Analysis Experiment Design

Sentiment analysis, as previously discussed in Section 2.4, Chapter II: Literature Review, is a task, where computation is performed on sentences to determine whether they express a positive, negative or neutral sentiment (Hardeniya et al., 2016). Sentiment analysis has been performed on the dataset, using the *Textblob* library, which function returns a tuple, representing the polarity and subjectivity of each individual user post, with sentiment indicated as a value nearer to 1 (i.e. positive) or nearer to -1 (i.e. negative). It is acknowledged that various algorithms, libraries and models can be used for more advanced sentiment analysis, however for the purpose of system prototyping, *Textblob* sentiment analysis is preferred due to its simplicity and speed of implementation (Srinivasa-Desikan, 2018). This type of method is categorised as lexicon-based sentiment analysis, with more advanced approaches being machine learning-based or hybrid (see Madhoushi et al., 2015).

In the build of a real-life application, this technique should be replaced with a more advanced model, such as Naïve Bayes, and even neural networks (Srinivasa-Desikan, 2018). In the current experimentation, the application of a Naïve Bayes model is also ensured, yet only for demonstration purposes, i.e. working with a small, self-defined semantic dictionary. To briefly summarise, the Bayes theorem is a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c), as demonstrated in Equation 3.3-2 (below), where:

- P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes);
- P(c) is the prior probability of class;
- P(x|c) is the likelihood which is the probability of predictor given class;
- P(x) is the prior probability of predictor (Rish, 2001).

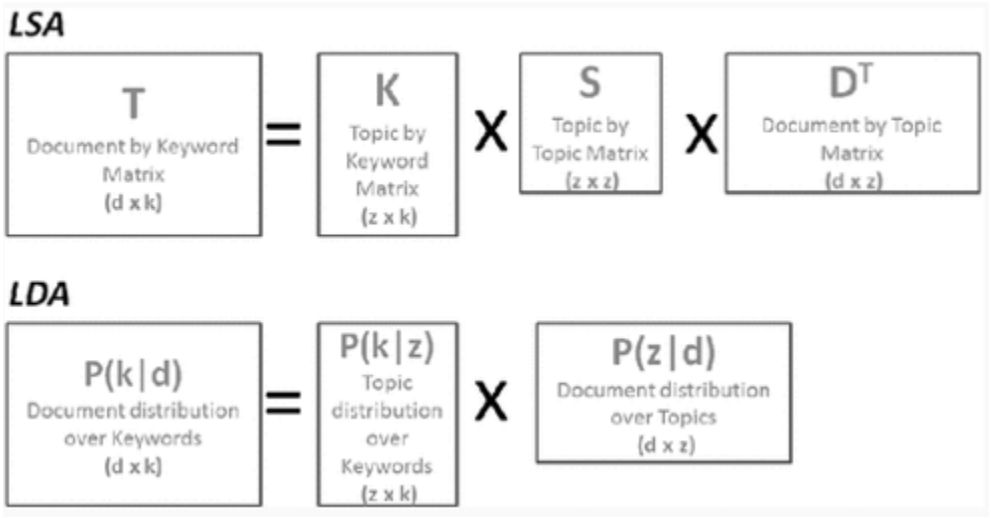
Equation 3.3-2 Naive Bayes theorem



3.3.2.5 Topic Modelling Experiment Design

In terms of topic modelling, considering the prevalence of LDA and LDA-hybrid methodologies in academic literature, demonstrated in Section 2.3.2, Chapter II: Literature Review, it has been chosen as one of the methodologies that will be assessed. Considering its underpinning assumption, i.e. that words in documents have underlying probabilistic distributions, which are used for topic discovery (Srinivasa-Desikan, 2018), another such model will be used for comparative evaluation – the LSA (Latent Semantic Allocation). LSA extracts and represents '*the contextual-usage meaning of words by statistical computations applied to a large corpus of documents'*, and is praised as '*a simple and efficient procedure for extracting topic representation of associations between terms from a term-document co-occurrence matrix*' (see Bergamaschi and Po, 2014: 252-3). Nonetheless, the model has been criticised for assuming a Gaussian distribution of the terms in the documents, which might not be the case for all documents, not handling non-linear document dependencies well, and utilising a Singular Value Decomposition (SVD), which is computationally intensive and hard to update as new data comes up (Joshi, 2018). These criticisms are shared by Bergamaschi and Po (2014), who analyse comparatively the LDA and LSA models on a movie review dataset. The current research builds on their experiments, as it provides an opportunity to test their performance on short-form, unstructured and significantly noisier data. Equation 3.3-3, below shows the matrix decomposition of both models

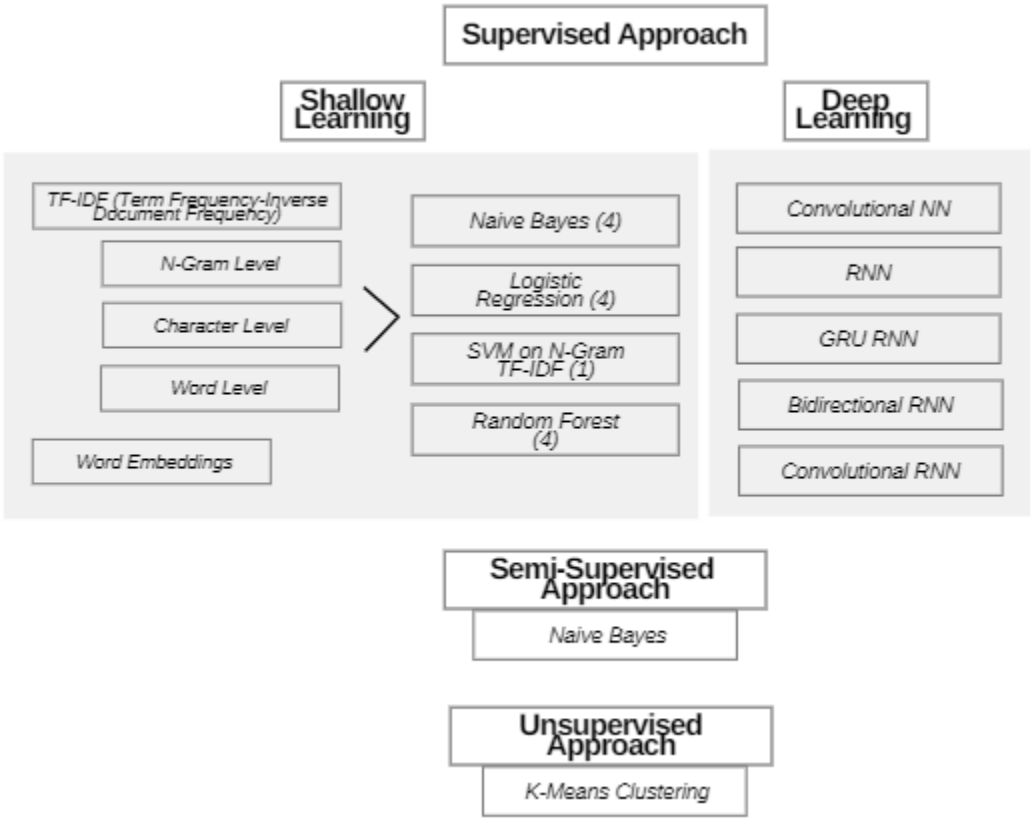
Equation 3.3-3 Matrix decomposition of LDA and LSA topic modelling techniques (Bergamaschi and Po, 2014)



3.3.2.6 Text Classification Experiment Design

In terms of text classification, **Error! Reference source not found.** (below) illustrates the models involved in the experiment. Specifically, text classification methods will be evaluated – (1) a supervised approach, using a shallow or deep learning methodologies, (2) a semi-supervised approach, using probabilistic classification and (3) an unsupervised approach, using a clustering algorithm. As evident from the figure, the advanced feature extraction explained previously will be used as part of the training for shallow learners, with each being trained on TF-IDF n-gram vectors, word level vectors, character level vectors, and word embeddings, with the exception of Support Vector Machines (SVM), which will only be tested on N-gram TF-IDF. Logistic regression, SVM and Naïve Bayes were chosen due to their prevalence in academic literature (see Gupta et al., 2017; dos Santos and Ladeira, 2014) in the context of classification. The development for deep learners will also be made as a result of the promise such models have demonstrated in previous academic research. This development will be completed and attached as Python code; however, its implementation and evaluation is out-of-scope for the current project, considering the lack of labels of the working dataset.

Figure 3.3-6 All text classification models, developed as part of the current research, arranged by approach type



Instead, prototyping will be done using the semi-supervised and non-supervised approaches demonstrated in **Error! Reference source not found.** (above). Naïve Bayes machine learning (introduced earlier, see Equation 3.3-2) has recently gained popularity in the context of text classification (Xu, 2018). It offers a semi-supervised probabilistic classification approach, based on a pre-defined dictionary, holding class descriptors. K-means clustering will also be applied on a test dataset to demonstrate the potential application of this model as well.

3.3.3 System Optimisation and Testing

3.3.3.1 Comparative Performance Evaluation Procedures

Machine learning experiments are typically criticised for testing on a few, pre-defined characteristics as opposed to using parametric (e.g. T-test and ANOVA) and non-parametric tests (e.g. Wilcoxon, Friedman, Quade) (Fernandez-Lozano et al., 2016; García et al., 2010). Statistical testing is challenging with lack of evaluation data and procedures when working with unlabelled data. Nonetheless, following the Fernandez-Lozano et al.'s (2016) critical analysis, the incorporation of external cross-valuation is integrated through human-agent evaluation of the system's performance, the protocol of which will be detailed below. Overall, however, the analysis

of results will be made in a qualitative manner, demonstrating storytelling through available data, collected from the performance of ML models and the performance of study participants on the same tasks as opposed to through statistical testing. The aim of the analysis is thus considered to be the evaluation of coherence of the results from various experimental procedures.

3.3.3.2 Potential for System Optimisation and Parameter Fine Tuning

There are some areas that have been identified as suitable for system optimisation and fine tuning. Firstly, when using a Naïve Bayes algorithm for sentiment classification, a more comprehensive semantic dictionary can be applied to ensure a more expert system. In terms of the lack of labels on data, Chollet (2018) recommends the use of self-supervised learning and *autoencoders*, which is something that can be explored as a continuation of this project. Transfer learning is another method that can be explored to improve the performance of classifiers, with many datasets and machine learning problems that exist suitable to provide a good foundation for the current system.

3.3.3.3 Comparative Performance Evaluation through Human-agents

Performance evaluation with participants will be performed through the distribution of an online survey, which will display texts and task users with the same goals as the machine learners, namely – to classify sentiment as positive, negative or neutral, identify the topics of the given text and perform classification into one of the customer journey classes. Surveys enable gathering the opinion of members of the public through direct communication, with participants being encouraged to answer questions in a truthful manner (Zikmund and Babin, 2012). The technique is cost- and time-efficient as it enables a quick sense-check of the system's output in a quantitative manner.

Table 3.3-3 (below) shows a summary of all associated methodological decisions that concern the use a survey-based data collection approach as part of academic research. The results from the survey and human-agent evaluation will be presented and discussed in-depth as part of Chapter IV.

Table 3.3-3 Methodological Desicions concerning survey experiment with human participants (a summary)

Design Aspect	***Choice***	***Rationale***
Ethical Approval	Received on 07.08.2019 from University of Strathclyde's Ethics Committee for CIS postgraduate research	N/A, detailed protocol attached as Appendix E
Research Protocol	Appendix F	N/A

Aim	Perform the same task as the system, on a smaller scale	Comparative Evaluation of Performance
Sampling Method	Selective/ Purposive (based on behavioural criteria), Convenience, Random	Statistical validation, Quantity of responses, Diverse pool of respondents, Variety of methods used to minimise bias (Collins, 2010; Dillman and Bowker, 2001; Tongco, 2007)
Recruitment	Social media Personal Network and Special Interest Networks, including ML, data analytics and others	Convenience, Speed, Efficiency (King et al., 2014; Rife et al., 2016)
Obtaining Permission	Electronic Consent Form	N/A, form attached as Appendix G
Question Types	Open-ended, Closed-ended	Ensuring opinion and expression are fully captured (Converse and Presser, 1986)
Survey Availability Period	08-11.08.2019	Research Scope, Time Availability
Software Used for Data collection	Qualtrics	Service Quality, Availability through University of Strathclyde
Software Used for Data Analysis	Excel, Qualtrics, Python Visualisation	N/A

3.3.3.4 Prototype Development

Following the concept validation and the relevant stages needed for system development and testing, a Graphical User Interface (GUI) prototype system design will be created in a digital platform to represent the client UI. Considering the growth of mobile internet users, discussed earlier in Chapter I, the most suitable mode of delivery of the created solution has been determined to be through a mobile application, which links with a corporate social media account and directly extracts sentiment, topics and performs text classification. The design of this system will be performed in Adobe Photoshop and XD, with visual appeal being an imperative, stemming from academic research suggestions that the visual appeal and UI quality of a prototype system can influence purchase decision of users (Wells et al., 2011; Chowdhury, 2019). Thus, a user-centred design approach has been adopted for the demo prototype development, which considers the client being a business organisation, looking to perform market analysis using social networks through opinion mining.

3.4 Limitations of Research

The primary experienced limitation throughout this research project was the lack of appropriate data, i.e. large in size and labelled for the tasks at hand. This is recognised in academic research as a common challenge in machine learning projects (Chollet, 2018, Bird et al., 2009). Although throughout the experimental procedures alternative systems were tested to demonstrate a potential system implementation, a manual labelling of data has not been attempted beyond a small sample of texts (100 data entries) due to this being a time-consuming task that is out-of-scope for the current project.

Some limitations stem from the data extraction methodology applied. Considering that keywords are used for data extraction, it is recognised that the dataset, used as part of the research is an imperfect representation of the consumer market in the given field. A further limitation is the restriction of data extraction from social networks as a result of API restrictions, which is a challenge commonly shared in the research field (Sapountzi and Psannis, 2018). Finally, it is recognised that the dataset although pre-processed still contains noise in the sense that not all texts are consumer generated (some are corporate tweets), and not all relate directly to B2C insurance.

A few limitations follow from the chosen research design. First, a lack of understanding exists of the links between topic models and text sentiments, which is shared in this type of research, as noted by Mei et al. (2007). Second, the limitation of time set the scope of development, resulting in a prototype of system, code and demonstration, with subsequent real-world implementation requiring a significant amount of development and system improvement. However, this research's aim being to find an optimal combination of existing models that can withstand comparative and participant evaluation tests, and simultaneously solve the given business problem. The rationale for doing so being the lack of a current method of solving this problem. It is believed that future research can address the system optimisation and fine-tuning recommendations made throughout this chapter and the analysis of results.

3.5 Ethical Considerations

The axiology, or otherwise the ethical considerations taken as part of research that requires human participants is a vital part for assessing the academic merit (Patton, 2002). The following paragraphs discuss the ethical concerns identified at the start of this research project, as well as how they have been handled.

For compliance with GDPR (2018) even social media user-generated data, extracted from a public repository requires permission to be obtained. This is ensured through extracting data from

public APIs, which enables the application to be classified as third-party, for which users have provided relevant access permissions as part of the platforms' terms and conditions.

The ethical treatment of participants in the system evaluation experiment must be ensured at all times throughout the experimentation period. To ensure this, ethical approval from the Ethics Committee in the University of Strathclyde has been obtained, for which a research protocol has been submitted, where all associated procedures and measures taken as part of the evaluation process are listed. Specifically, as part of the survey participants have been informed of their rights in relation to the experiment, and also given the option to withdraw their participation at any time prior to submission of their completed survey answers.

Participant anonymity has kept throughout, with no personal or identifiable data being collected or stored, which ensures compliance with GDPR (General Data Protection Regulation) (2018), ESRC's (Economic and Social Research Council) ethical guidelines of 2015, as well as with academic research guidance for conducting online surveys (see Cooper and Schindler, 2014; Johnson and Rowlands, 2012; Flick, 2014) and for using social media data for data analytics system development (see Taylor and Pagliari, 2018). In order to assist the evaluation of research merit, however, participants have been instructed that their system evaluation sheets and all other project-related data, generated from the survey will be kept and published as part of this project's completion.

3.6 **Evaluation of Academic Rigour**

3.6.1 Replicability

Replicability of the current research is ensured through providing supporting documentation that illustrates the processes followed at all stages of the research, specifically a summary of research strategy for the literature review (Appendix C), a summary of the data extraction procedures, search terms and APIs (see Appendix D), associated code for model development (submitted with thesis), the protocol followed for human-agent performance evaluation and the survey created with associated response data (see Appendices F and H, respectively).

Running multiple tests allows categorisation of observations and limits the risk of factorial dependency (Japkowicz and Shah, 2011). Reporting on settings ensures that the experiment can be conducted again for external validation. Using the same 'random seed' (=122) during training is another method of ensuring replicability. Both have been considered throughout all development and experimentation procedures.

3.6.2 Reliability and Triangulation

Reliability measures whether the study's results can be repeated in another environment (Yin, 2003; Mason, 2002). Triangulation is defined as cross-use of two or more independent sources of data or data handling approaches, which are used to corroborate research findings within a given study (Stebbins, 2001; Buchanan and Bryman, 2009; Saunders et al., 2016; Bell et al., 2018). Testing using a mixed-method approach ensures reporting of results, which are validated through comparative analysis. The methodologies chosen for experimentation are selected in a manner that provides opportunity for triangulating results from the current experiments with other academic literature.

3.6.3 Validity and Generalisation

Validity measures the extent to which the research question is addressed into the methodology (Lewis and Ritchie, 2003; Mason, 2002). Validity of results will be ensured through user cross validation (Fernandez-Lozano et al., 2016). Moreover, a constant comparative method is used throughout the analysis of results process, which allows for any inconsistencies to be brought to the reader's attention (Lewis and Ritchie, 2003; Glaser and Strauss, 1967).

In order to ensure model generalisability (otherwise referred to as external validity), the classification problem will be shaped to minimise the probability of mis-classification errors during training, allowing minimisation of the potential of overfitting the given data (Japkowicz and Shah, 2011). A common concern is the lack of system application in external domains, however, the system designed as part of this research can be utilised for solving a variety of real-life business problems in a number of domains. Please refer to Chapter V: Conclusion and Recommendations where the future research opportunities are discussed in greater detail.

3.7 **Conclusion**

In short, this chapter has detailed the methodology used to answer the research questions. The discussion delivered both a holistic research overview with rationale for relevant choices, as well as a detailed explanation of all associated techniques and procedures, the limitations of study, ethical considerations and methods used to ensure academic rigour. Next, follows the Analysis Chapter, where the results from all experimental procedures will be presented and discussed in light of the current and other academic research.

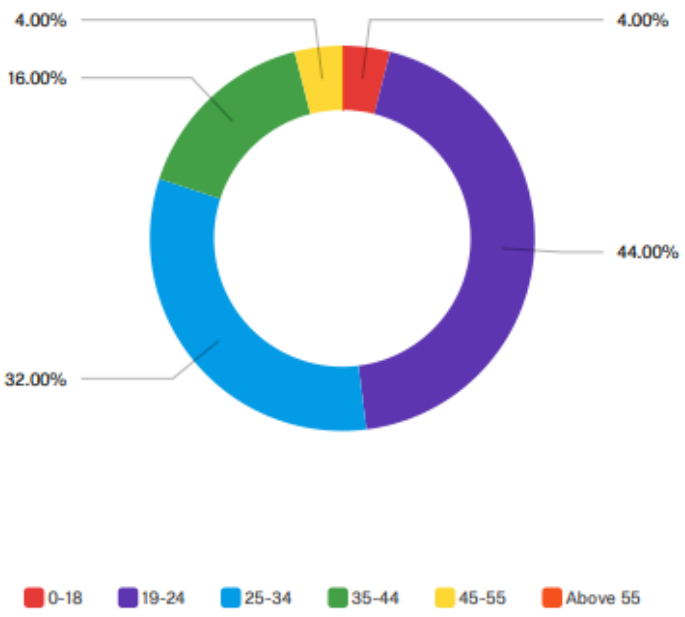
4. CHAPTER IV: ANALYSIS

4.1 Introduction

The following chapter will present the results from all conducted experiments and provide analysis and interpretation of the findings. The insight will be linked with previously posed research questions in short discussions throughout the chapter, where various components of the system are assessed.

Following the distribution of a survey, a total of 58 responses were recorded, some of which upon evaluation appeared to be partial. Nonetheless, the data from all entries was analysed per individual question. All responses, alongside the questions, which participants were asked are linked as Appendix H. Figure 4.1-1, below shows the age distribution of the participants in the survey, whereas Appendix I demonstrates a location map of respondents, who took part in a browser mode that enables location tracking. The collected demographic data demonstrates a good age distribution, with no majoritarian group, whereas the location data shows evidence of respondents from the UK, Germany, US, Bulgaria, Poland and Czech Republic.

Figure 4.1-1 Age Distribution of Survey Participants



The chapter is organised as follows: first, sentiment analysis experiment results are discussed in Section 4.2, where data from model development and survey responses are collectively discussed, followed by topic modelling and text classification results, in Sections 4.3. and 4.4, respectively. Each of those sections will present results from technical performance evaluation,

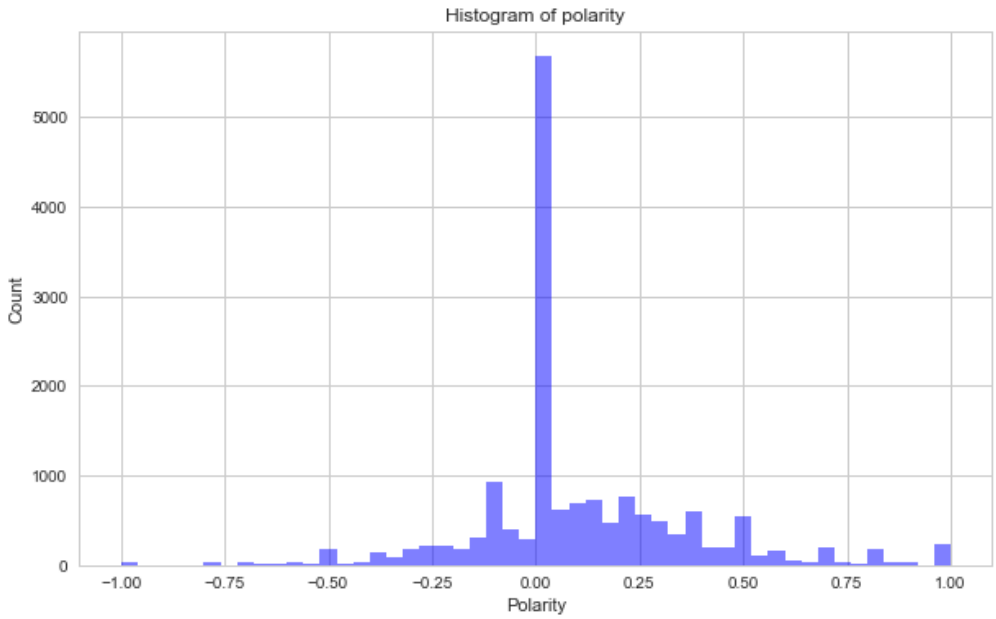
as well as from the human-agent performance evaluation. Important insights and challenges will be highlighted throughout.

4.2 Sentiment Analysis Experiment Results

4.2.1 Presentation of Results

Considering the lack of sentiment analysis labels, the first task of the experiment was to use a textblob library to extract sentiment. The histogram of sentiment polarity attached as Figure 4.2-1 (below) demonstrates that just over a third of texts express a neutral sentiment polarity.

Figure 4.2-1 Histogram of Sentiment Polarity, extracted from Textblob sentiment classification



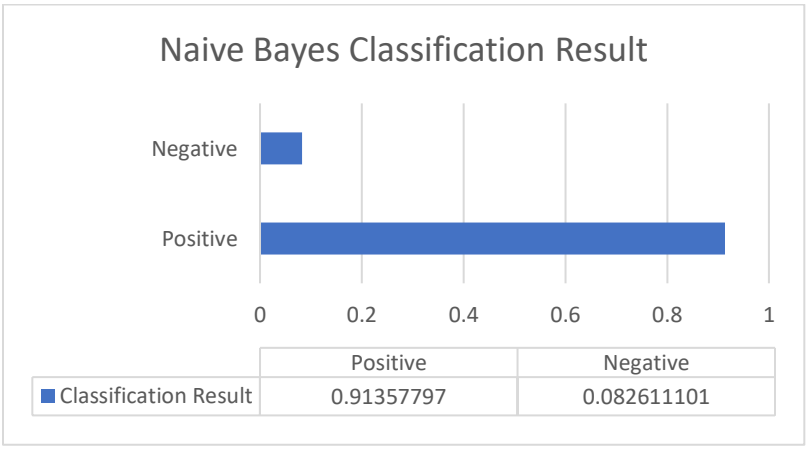
For further visualisation of the performance of text blob sentiment analysis, the table below is attached (Table 4.2-1), which shows generated word clouds from term frequency words from texts that have been labelled with extreme negative or positive sentiment (-1 or 1, respectively). Considering that the size of the words in these visualisations corresponds with the frequency of word use and the histogram in the previous Figure 4.1-1 shows that comparatively extreme positive texts are more than extreme negative texts, it is interesting to note that there is a greater vocabulary consistency where negative sentiment is expressed. An interesting characteristic is the use of negative sentiment-intense adjectives, e.g. terrible, worst, disgusting, insane, pathetic. Likewise, some of the key terms according to the size in this visualisation shows the prevalence of texts mentioning a company, claim, car, service, money and people. Contrastingly, company, car and life policy do appear to be more frequently used than other positive words in the extreme positive sample, however in comparison to the negative word cloud their importance is lesser.

Table 4.2-1 Word cloud with negative (left) and positive (right) sentiment polarity, extracted from textblob classification

Wordcloud with most frequent words from texts with negative sentiment	*Wordcloud with most frequent words from texts with positive sentiment*
	

To demonstrate an alternative means of extracting sentiment, a prototype Naïve Bayes algorithm was applied with a self-defined dictionary provided the following result, however it can be argued that this can be improved significantly through integration of external semantic dictionaries. Contrastingly to the textblob evaluation of semantic features, this model classified the majority of texts as positive (see Figure 4.2-2, below). Furthermore, in comparison to the textblob distribution, the Naïve Bayes algorithm classified a similar number of texts as negative, namely 1,344 texts from a total of 16,269 in comparison with just under 2,000 text for textblob.

Figure 4.2-2 Naive Bayes sentiment classification result

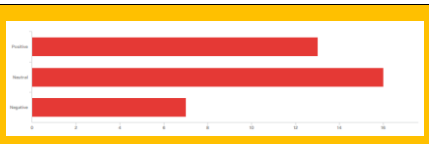
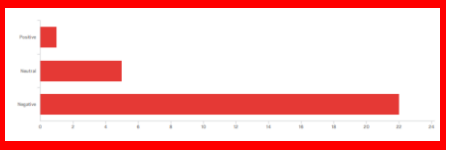


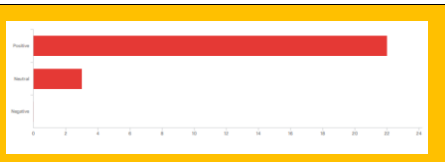
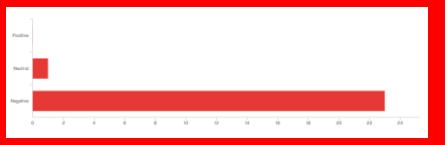
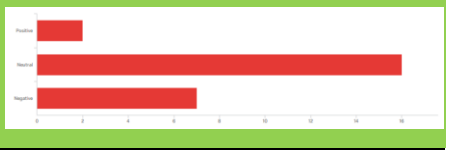
4.2.2 Comparative Analysis of Automatic and manual sentiment Classification

The following Table 4.2-2 presents comparative analysis of automated sentiment analysis using the textblob library versus sentiment analysis from the survey participants on the same texts. The automatic and manual sentiment analysis of these texts agreed 4 out of 9 times. Wherever there

was a disagreement, a significant error was made by the automatic analysis from a business standpoint in example texts 3 and 8 (in red), which demonstrates the necessity to implement system fine-tuning or an alternative, more advanced method. The remaining 3 cases (i.e. where there was a disagreement, and the automatic classifier was not severely wrong; texts 1, 5, 7 (yellow colour)), demonstrate in two of the cases (text 1 and 5) that the survey participants show a less unified response in their classification data.

Table 4.2-2 Comparative Analysis of results from manual and automated (lexicon-based) sentiment classification on selected user-generated texts

User-generated text	*TextBlob Classification*	*Human-agent Classification*	*# of Responses*	*Response distribution (top = positive, middle = neutral, bottom = negative)*
In the worst case of flooding, I hope to get a She Shed with my insurance money. #HurricaneBarry2019 #HurricaneBarry	Positive (=1)	Neutral	36	
@COMPANY Worst insurance company I have ever seen. As per my experience they don't provide the offered insurance amount . It's a trap for the customers. They just loot the people. I don't get how @USER has allowed such companies to operate their business	Negative (=-1)	Negative	32	
@COMPANY @USER We are trying our best to get a life. To get a roof-to get walls-to get the insurance to return our calls. All the while being told to 'get over it'.	Positive (=1)	Negative	28	
Check out the easiest and quickest way to find affordable coverage with us. It only takes a few minutes to compare the best quotes from a variety of providers, giving you the most choice when it comes to finding your home insurance policy. [link]	Positive (=1)	Positive	26	
@USER Thankfully her kids aren't in school yet and the insurance company finally gave her a rental. But yeah, the complication of it all is a pain in the ass. And she's JUST back.	Neutral (=0)	Negative	27	
Does @COMPANY have an accident policy? I know @COMPANY does and you are insured as a rider. As a rider with @COMPANY, are you insured in the case of an accident?	Neutral (=0)	Neutral	25	

The wife got her dream car today (aside from a G-Wagon) Insured by @COMPANY #BIGCoverage	Neutral (=0)	Positive	25	
@COMPANY avoid at all costs. This place is a joke. Your insured member was at fault and caused an accident involving 3 other cars two weeks ago. Countless calls and nobody has returned my call regarding my damaged vehicle. Your claims adjuster will not return calls	Neutral (=0)	Negative	24	
@USER A feedlot I work with just told me yesterday insurance won't allow them to put plastic on silage pile this fall because of worker safety	Neutral (=0)	Neutral	25	

4.2.3 Discussion of results

When comparing the findings with those of other researchers, a rationale for inconsistencies can be provided in the degree of subjective interpretation of texts, which can skew both study participants and sentiment-based approaches' results (Rosenthal et al., 2015). Language ambiguity is a challenge, especially relevant for short-form texts, as affirmed by academic research (Chen et al., 2011; Rao et al., 2016; Ittoo et al., 2016). The textblob library approach also fails to capture semantic relationship of words, further hindering performance (Sriram et al., 2010; Tang et al., 2019). As discussed previously in Chapter III, this model is useful for prototyping purposes only, yet has produced good performance considering the implementation speed and ease.

4.3 **Topic Modelling Model Evaluation**

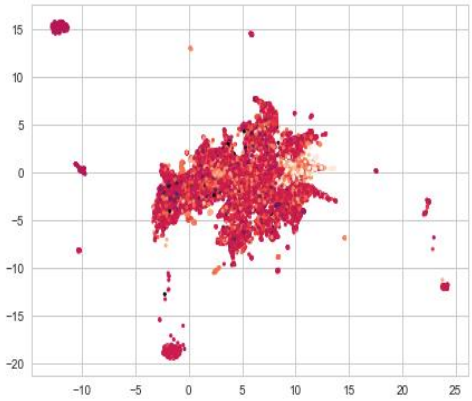
4.3.1 Presentation of Results

With LSA topic modelling, which uses SVD – a matrix factorization method, which represents a matrix as a product of two matrices, 19 topics were extracted with the key words contributing to each one illustrated in Table 4.3-1, below. Mathematic evaluation of this model is not currently supported in Python, and interpretation of coherence arguably requires expert knowledge is needed in the domain. Nonetheless, from a business marketing point of view, it can be argued that these topics affirm that consumers would seek advice regarding their insurance cover (e.g. topics 6, 8, 12, 13, 14), with price ('cheap' (topics 17, 19), 'money' (topic 5)) being a commonly featured term. The results also demonstrate that cover is often discussed online, as well as life

and health insurance being commonly featured. In terms of speed, this technique is faster as opposed to LDA topic modelling.

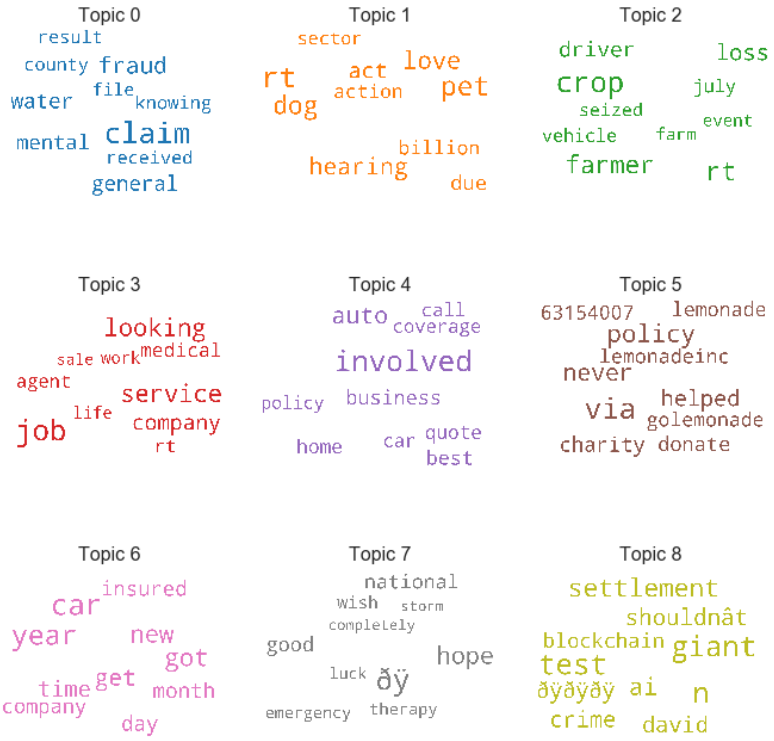
Table 4.3-1 LSA topic modelling results

Topic #	*LSA*	*LSA Topic-Sentiment (Each dot represents a single user-generated text and the colours represent the sentiment polarity)*
1	Umbrella limit liable whats asset involved putting	
2	life company policy health insured need httpstco	
3	life policy farm life insurance family issued protect	
4	insured life fully licensed money time drive	
5	company life insured people money year like	
6	health cover care plan people need provider	
7	business health life insured care small seeking	
8	cover business policy need like know dont	
9	looking like work good year need time	
10	httpstco looking cover work medical agent sale	
11	best cover httpstco looking recommend good need	
12	need life company help lemonade cover know	
13	need httpstco policy know dont year provider	
14	policy looking need insured health company claim	
15	year need claim customer state provider farm	
16	good recommend need service agent claim time	
17	quote money time free online need cheap	
18	time money auto make claim best know	
19	Know year don't good quote cheap health	



Comparatively, the LDA algorithm, due to its prevalence allows a more comprehensive performance analysis, enabled through established tools for the purpose. Figure 4.3-1 LDA topic models (below) shows the topic models extracted through LDA in the form of word clouds, where the size corresponds with the term's weight. It can be argued that these topics offer a comparatively more comprehensive idea of the various topics that are extracted, with words being pared accordingly, e.g. in topic 2 – farmer, farm, crop, in topic 4 - auto, coverage, car, quote.

Figure 4.3-1 LDA topic models (Word Cloud Representation)



For the LDA model, the perplexity score and topic coherence score were used to judge the efficiency of the model, illustrated in Table 4.3-2, below. The perplexity measures the marginal likelihood on a held-out test set (Yan et al., 2013), whereas the topic coherence score is an automated measure for evaluation of topic quality (Mimno et al., 2011). As part of system optimisation for system implementation, the coherence score can be re-examined through comparative evaluation in relation to the number of LDA topics to find the optimal score.

Table 4.3-2 Evaluation Metrics for LDA performance

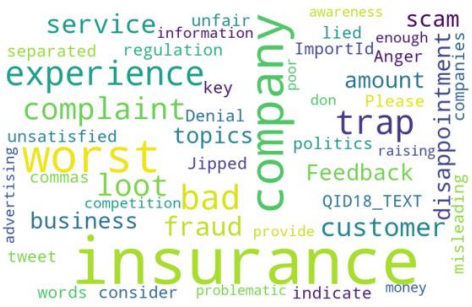
Measure	*Score*
Perplexity Score	-12.883395715070044
Coherence Score	0.4485013880049079

Various visualisation options are possible for LDA, most notably - the *pyLDAvis* library, which allows once the models is trained for users to process it dynamically, accessible through a web browser or Jupyter notebook, identifying the key topics for each text and their prevalent words (listed in the submitted code). A visualisation is also presented in Appendix J to show the LDA topic weights, descriptors and topic-term probability chart for various tweets.

4.3.2 Human-Agent Performance Evaluation Results

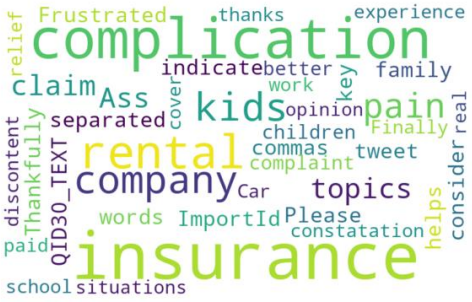
Finally, the topics that were picked by study participants will be presented for comparative evaluation. Table 4.3-3 presents the synthesised topics from response submissions, alongside the text that users were asked to evaluate. A significant difference in the topic models generated automatically as opposed to those generated by study participants is the interpretation of text, incorporated in the latter. For example, automatic topic modelling extracts topics from words already existing in the dataset, whereas participants interpreted the texts, e.g. the first text in Table 4.3-3 resulted in generation of topic such as ‘disaster’, ‘fearful’, ‘attention’, ‘humour’, and some of the topics for the second text included ‘unsatisfied’, ‘disappointment’ or ‘anger’. It can be concluded from the comparison that manual topic generation is interpretative and qualitative, whereas the topic models tested in the experiment offer a probabilistic, quantitative overview of the text.

Table 4.3-3 Manually generated topic models by study participants for selected texts

User-generated Text	*Word cloud generated from human-agent topic entries*
In the worst case of flooding, I hope to get a She Shed with my insurance money. #HurricaneBarry2019 #HurricaneBarry	
@COMPANY Worst insurance company I have ever seen. As per my experience they don't provide the offered insurance amount . It's a trap for the customers. They just loot the people. I don't get how @USER has allowed such companies to operate their business	

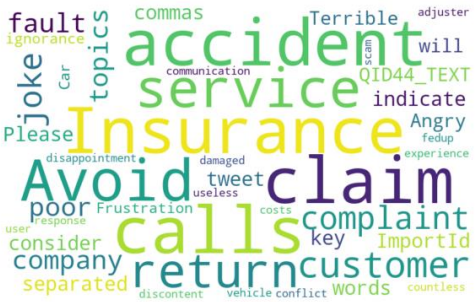
@COMPANY @USER We are trying our best to get a life. To get a roof-to get walls-to get the insurance to return our calls. All the while being told to 'get over it'.

Check out the easiest and quickest way to find affordable coverage with us. It only takes a few minutes to compare the best quotes from a variety of providers, giving you the most choice when it comes to finding your home insurance policy. [link]



@USER Thankfully her kids aren't in school yet and the insurance company finally gave her a rental. But yeah, the complication of it all is a pain in the ass. And she's JUST back.

Does @COMPANY have an accident policy? I know @COMPANY does and you are insured as a rider. As a rider with @COMPANY, are you insured in the case of an accident?

The wife got her dream car today (aside from a G-Wagon) Insured by @COMPANY #BIGCoverage	
@COMPANY avoid at all costs. This place is a joke. Your insured member was at fault and caused an accident involving 3 other cars two weeks ago. Countless calls and nobody has returned my call regarding my damaged vehicle. Your claims adjuster will not return calls	
@USER A feedlot I work with just told me yesterday insurance won't allow them to put plastic on silage pile this fall because of worker safety	

4.3.3 Discussion

Previous studies that compare the performance of LDA and LSA consistently affirm the superiority of LSA/SVD, especially when compared with human judgement (Stevens et al., 2012; Bergamaschi, S. and Po, 2014). When evaluated by users, the LDA model fails to achieve good performance (Bergamaschi, S. and Po, 2014), on the basis of whether their topic recommendation was similar or not similar to the machine's. The current study's experiment has not produced sufficient data to evaluate the superiority of one opposed to the other, however, the LDA model was comparatively slower in terms of processing time.

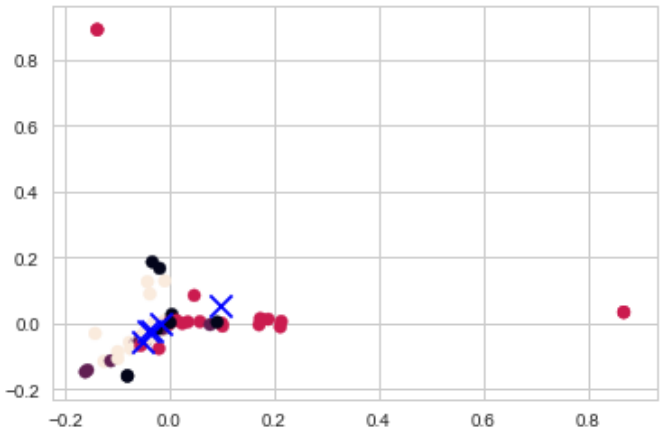
4.4 Text Classification Model Evaluation

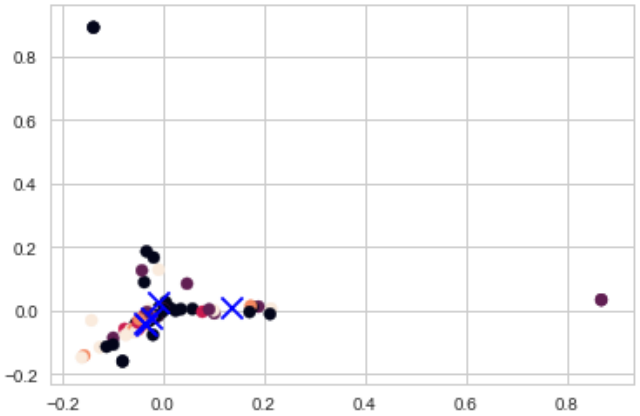
4.4.1 Presentation of Results

In order to test the performance on the supervised models, detailed in the Methodology chapter, a test mini dataset has been prepared, which contains 100 data entries with relevant labels. This was done to demonstrate the possibility of running all models, however considering the size of this small dataset no results were produced from the shallow and deep supervised learners. Nonetheless, the development for this section of the experiment is attached in the code and is ready for comparative evaluation in future projects with different data or labelled data.

Alternative solutions have been sought in semi-supervised and unsupervised models, which typically require no labelled data until the evaluation stage. Specifically, K-means clustering and minibatch K-Means have been cross-compared on the mini-dataset (with 100 samples), trained to recognise 6 clusters (5 - for the customer journey stages and 1 - for unrelated texts). Both were also tested on the full dataset, however failed to yield results due to memory issues, even when running on Python 64bit. Table 4.4-1 (below) shows the cluster matrices, as well as the homogeneity scores for the models.

Table 4.4-1 K-Means clustering text classification performance

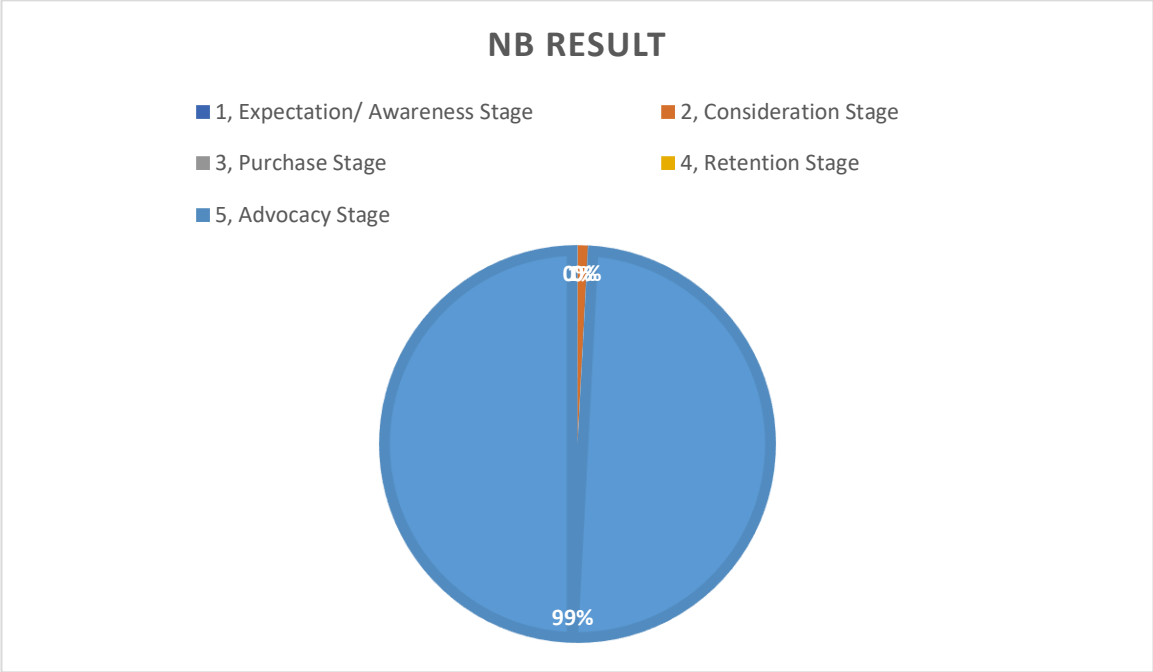
Specifications	*Matrix with Topic clusters*
	Minibatch K-Means
Trained on 100 labelled samples *6 clusters* *Homogeneity score 0.1298367891 3185587*	
	KMeans

Trained on 100 labelled samples 6 clusters *Homogeneity score* *0.1298367891 3185587*	

A semi-supervised dictionary-based approach was applied that uses Naïve Bayes with a self-defined dictionary. The dictionary and model results are listed in Table 4.4-2, below. The dictionary is a prototype model, which can be amended to better suits the needs of the classification, and the results are based on word-level analysis, yet following implementation, it is considered more suitable to apply sentence-level or document (i.e. short-form text level) implementation.

Table 4.4-2 Word-level Dictionary-based Text Classification with Naive Bayes Results

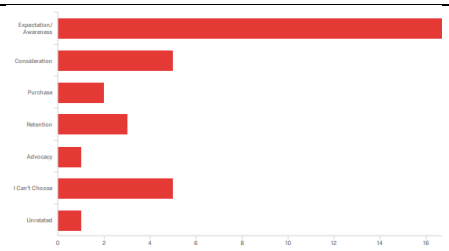
Stage of the Customer Journey	*Specifications/ Dictionary*	*Result*
1, Expectation/ Awareness Stage	'what', 'when', 'need'	0.00018439977872026554
2, Consideration Stage	'recommend', 'looking for', 'can you', 'compare', 'considering'	0.007990657077878173
3, Purchase Stage	‘just bought', 'just started', 'started', 'new policy', 'new car', ‘buying’	0.0
4, Retention Stage	'update policy', 'my new policy', 'repurchase', 'buy again', 'new policy'	0.9669309730161657
5, Advocacy Stage	'bad', 'terrible','useless', 'hate', ':(', 'dissappointed', 'avoid', 'love', 'company', 'price', 'service'	0.003503595795685045
Visualisation of result		

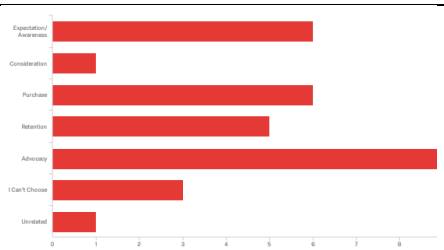
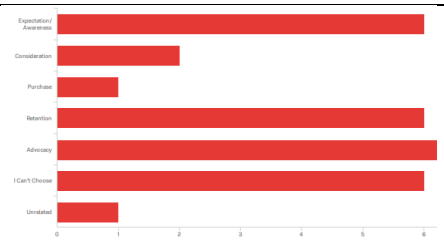
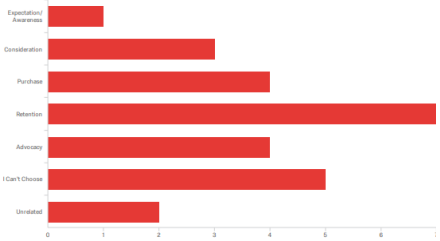
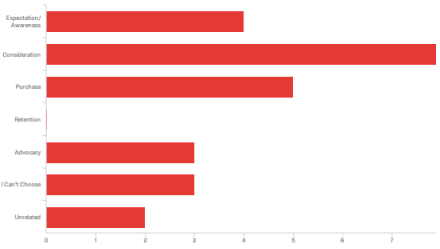
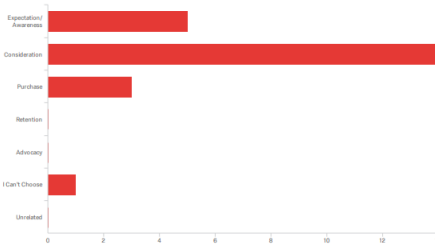
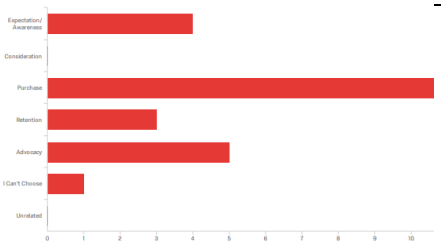


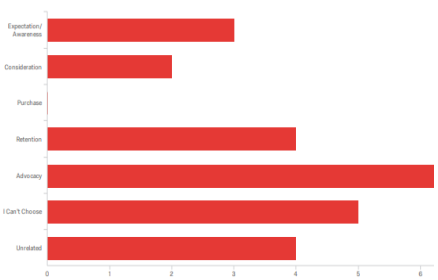
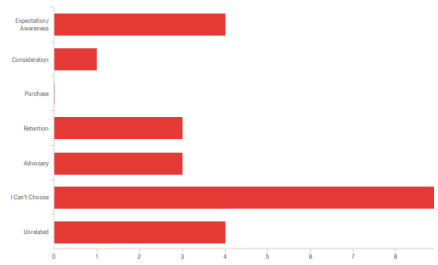
4.4.2 Human-Agent Performance Evaluation Results

The results from the text classification with study participants demonstrate the difficulty of this task without prior expert knowledge. Table 4.4-3 (below) shows the extracted user-generated text, and the classification given by participators, as well as the distribution and response number. Only for two texts there is a majority that agrees on the correct class that should be assigned and, in both cases, it is an insignificant majority. As demonstrated by the sample picked for human-agent classification, there is a variety of texts chosen for the users to apply each class at least once, including the 'Unrelated' (6th class, for the last text). In the two cases where the majority of respondents were in agreement, one of the classes agreed upon cannot be considered correct, as the consumer that has written the tweet has already purchased insurance, so a purchase or retention class might have been more suitable.

Table 4.4-3 Results from manually generated by study participants text classification on selected texts

User-generated Text	*Customer Journey Classification*	*Distribution*	*# of Responses*
In the worst case of flooding, I hope to get a She Shed with my insurance money. #HurricaneBarry2019 #HurricaneBarry		**Expectation/ Awareness 52.78%** Consideration 13.89% Purchase 5.56% Retention 8.33% Advocacy 2.78% I Can't Choose 13.89% Unrelated 2.78%	36

@COMPANY Worst insurance company I have ever seen. As per my experience they don't provide the offered insurance amount . It's a trap for the customers. They just loot the people. I don't get how @USER has allowed such companies to operate their business		Expectation/ Awareness 18.75% Consideration 3.13% Purchase 18.75% Retention 15.63% **Advocacy 31.25%** I Can't Choose 9.38% Unrelated 3.13%	32
@COMPANY @USER We are trying our best to get a life. To get a roof-to get walls-to get the insurance to return our calls. All the while being told to 'get over it'.		Expectation/ Awareness 20.69% Consideration 6.90% Purchase 3.45% Retention 20.69% **Advocacy 24.14%** I Can't Choose 20.69% Unrelated 3.45%	29
Check out the easiest and quickest way to find affordable coverage with us. It only takes a few minutes to compare the best quotes from a variety of providers, giving you the most choice when it comes to finding your home insurance policy. [link]		Expectation/ Awareness 15.38% **Consideration 34.62%** Purchase 19.23% Retention 0.00% Advocacy 11.54% I Can't Choose 11.54% Unrelated 7.69%	26
@USER Thankfully her kids aren't in school yet and the insurance company finally gave her a rental. But yeah, the complication of it all is a pain in the ass. And she's JUST back.		Expectation/ Awareness 3.70% Consideration 11.11% Purchase 14.81% **Retention 29.63%** Advocacy 14.81% I Can't Choose 18.52% Unrelated 7.41%	27
Does @COMPANY have an accident policy? I know @COMPANY does and you are insured as a rider. As a rider with @COMPANY, are you insured in the case of an accident?		Expectation/ Awareness 20.00% **Consideration 64.00%** Purchase 12.00% Retention 0.00% Advocacy 0.00% I Can't Choose 4.00% Unrelated 0.00%	25
The wife got her dream car today (aside from a G-Wagon) Insured by @COMPANY #BIGCoverage		Expectation/ Awareness 16.00% Consideration 0.00% **Purchase 48.00%** Retention 12.00% Advocacy 20.00% I Can't Choose 4.00% Unrelated 0.00%	25

@COMPANY avoid at all costs. This place is a joke. Your insured member was at fault and caused an accident involving 3 other cars two weeks ago. Countless calls and nobody has returned my call regarding my damaged vehicle. Your claims adjuster will not return calls		Expectation/ Awareness 12.00% Consideration 8.00% Purchase 0.00% Retention 16.00% **Advocacy 28.00%** I Can't Choose 20.00% Unrelated 16.00%	25
@USER A feedlot I work with just told me yesterday insurance won't allow them to put plastic on silage pile this fall because of worker safety		Expectation/ Awareness 16.00% Consideration 4.00% Purchase 0.00% Retention 12.00% Advocacy 12.00% **I Can't Choose 40.00%** Unrelated 16.00%	25

4.4.3 Discussion

Considering the lack of experimental data of shallow and deep supervised learners, a brief discussion is provided of previous studies that have applied the methods referenced earlier in Section 3.3.2.6, in the Methodology Chapter, to help identify the reasons of failure to utilise these models in the present process. Through a comparative analysis of SVM, Naïve Bayes and probabilistic models for sentiment extraction (LDA, LSA), Song et al.'s (2014) survey concludes that text classification at present can be achieved through a semi-supervised approach, with performance being boosted by ensemble techniques. Zhang et al.'s (2015) study evaluates the performance of character-level CNNs on a several heavily-populated, large datasets against shallow learner approaches, all using advanced feature extraction as detailed in the present piece, showing promise in the model, yet demonstrating the importance of input data for such a complex architecture. Bidirectional RNNs trained on feature vector representations demonstrate promising results for text classification as well, however the application in academic research benefits from six large, multi-variant datasets (see Zhou et al., 2016). Other studies have proposed utilising the attributes of semantic analysis and word embeddings to improve the performance of deep learning techniques (Wang et al., 2016). Overall, the triangulation with previous literature demonstrates that to perform text classification with a deep learning approach a variety of experimental procedures are required, with testing and training being performed on a large dataset.

The results from the current study demonstrate the difficulty in extracting sufficient contextual information from the short-form text to successfully classify the text into a behavioural category.

The survey with study participants affirmed the complexity of the task without expert knowledge of the customer journey concept and without applying deep logical reasoning. Nonetheless, there are a number of insights and opportunities for future research, stemming from the text classification experiments, which will be discussed in Chapter V.

4.5 **Conclusion**

To sum up, this chapter presented results from the comparative analysis and evaluation of two sentiment models – an extraction algorithm and a classifier, two probabilistic topic modelling techniques and two text classification/clustering techniques. The models' performance was cross-referenced with the responses of study participants, who attempted to perform the same tasks as the machine learning algorithms, as well as was triangulated with data from previous academic research. The next chapter will discuss the findings in the context of the research questions and aim, presenting a wider discussion of the implications of these findings and future research opportunities.

5. CHAPTER V: CONCLUSION AND RECOMMENDATIONS

5.1 **Introduction**

The following chapter will present a short summary of the problem statement, project findings, and present applications of those in the form of recommendations. The implications of this research for professionals, working in the field of textual analytics will be provided, as well as for academic researchers. Arguably, in consideration of the project's problem statement and resulting prototype system, numerous research opportunity stems were identified, which will be elaborated on in Section 5.5. Finally, a reflection statement will be provided to summarise the project experience from a researcher standpoint.

5.2 **Recap of Problem Statement**

In Section 1.3, Chapter I, the aim of the study was identified as the creation of "*a system of tools that can extract topics and associated sentiment polarity from social media data, and subsequently allocate user-generated text in pre-defined classes that correspond with the stages of a purchase customer journey"*. The drivers of this problem were identified to be (1) the availability of data on social media that can be transformed to insight for organisations; (2) the growth of NLP research as a discipline, which has resulted in the optimisation of a plethora of models for sentiment analysis, topic modelling and text classification, which can be utilised by organisations to reduce or eliminate the manual completion of these tasks in a scalable manner.

The potential benefits of automation of the examined processes can be re-affirmed following the conducted experiment, especially considering the ease of application of final algorithms and the results produced with relatively low-maintenance processing capacity and data cleaning. To recap from Section 1.8, such benefits for organisations include:

- Informed planning of business operational goals;
- Capacity to prioritise areas, identified as problematic;
- Improvements in targeted responsiveness;
- Improved communication with consumers and 'feel for the market';
- Ability to monitor consumers and their responses to stimuli intelligently and holistically;
- Proactive responsiveness to identifying the topics;
- Strategic planning support;
- Potential to improvement of digital marketing content strategy;
- Ability to target market micro-segments with marketing communication or promotional activities;
- Potential of improving companies' relationship marketing efforts;

- Potential of improved customer retention;
- Capacity to identify and understand customer journey process ‘leaks’ and their causes.

5.3 Key Findings and Associated Conclusions

In terms of the performed sentiment analysis, there were some identified inconsistencies when comparing the automated approach with the manual classification process, with a couple of significant errors identified. However, when considering the lack of approach adaptation to the dataset and fine-tuning prior to implementation, it can be argued that the automation approach performed okay. The implementation of the automated approach was time-efficient and non-exhaustive effort-wise. This suggests that an automated approach, which benefits of invested time in terms of fine tuning, and experimentation on the dataset can substantially out-perform manual sentiment classification.

Topic modelling experimentation demonstrated engaging results from both a technical and manual standpoint. Although the superiority of neither of the compared models can be affirmed due to lack of parametric and non-parametric tests, for the purpose of the proposed system the LDA is chosen as the superior model, primarily due to its ease and speed of implementation, topic coherence and associated topic visualisation tools that are available for developers, using this model, such as pyLDAvis. Manual topic modelling arguably demonstrated what the compared models lack – logistical interpretation of text. The conclusion that dimensionality reduction methods (i.e. LDA) are flawed compared to human text interpretation is not new (see Mimno et al., 2011), nor that interpretation of topic model coherence is subjective (Chang et al., 2009), which was identified as being a difficulty in the topic model experiment results evaluation sections; yet the reoccurrence of these problems serves as a rehash of research gaps, to which no tested solutions exist, opening opportunities for future development of logic-based topic modelling instruments.

The text classification group of experiments proved being the most difficult of the three procedures. Although findings alternative solutions to supervised learning approaches is not difficult, it is recognised that the classification accuracy achieved by the proposed methods is not sufficient for industry application prior to fine-tuning of the model. Contrary to expectations, the type of data did not hinder the progression of the experiment as did the lack of labels. This demonstrates the need for more academic research that shows improvements on the processes of working with real-time and unlabelled texts, as opposed to offline, with labelled and publicly recognised datasets. It is believed that this can help advance NLP through demonstrating solutions for real business challenges, as opposed to creating superior models (e.g. deep learning), to which researchers have access to and the programming knowledge to develop (as

demonstrated in the current research), but lack the opportunity to implement due to restrictions, caused by irregularities in data.

5.4 Recommendations and Implications for key Stakeholders

5.4.1 Presentation of Demo System

A prototype system design is illustrated in below. It shows three screens: (1) Topic Search, where the user can enter the topics of interest to them, (2) Topic-Sentiment Visualisation, where the topics are plotted with circles in respect of their size on a two-dimensional plot that represent sentiment polarity, and (3) Customer journey, which is a screen that is unveiled in the consumer taps on one of the topics. The customer journey screen breaks down the stages of the customer journey consumers have been identified to be in, based on the contextual information of the texts they have posted on social media for the selected topic.

Figure 5.4-1 Demo Mobile App Functionality Prototype

5.4.2 Process Automation

Although in a prototype format, the developed system demonstrates the availability of sentiment polarity classification and topic modelling algorithms that are easy to implement and coherent as a minimum viable product for social media textual analytics. Considering the previously affirmed

potential benefits of implementation of a text insight extraction, the lack of action will inevitably result in loss of organisational competitive advantage in a fast-paced data-driven market.

5.4.3 Working with real-time, unlabelled, short-form data

A recommendation for academics involves seeking ways to design and implement the advances of machine learning in a non-experimental manner, outwit the controlled conditions that labelled data provides. A way to translate this is to use models that are pre-trained and utilise transfer learning for testing on unlabelled datasets. Although staggering progress has been made following recent developments in deep learning, consultations with industry professionals in textual analytics demonstrate that such models are out of reach for implementation in small and medium-size organisations. A development in the democratisation of this knowledge is increasing its accessibility, which as demonstrated by the current research might be hindered by the lack of labelled data.

5.5 **Future Research Opportunities**

5.5.1 Application in Learning Analytics and Education Enhancement for University of Strathclyde

In Section 3.2.2 it was explained that user-generated texts will be considered as data points alone, with user stories remaining unexplored and unaddressed. This research paradigm enables to think about the current research and future research opportunities in a holistic manner, extracting the fundamentals of research and applying them in other contexts.

The University of Strathclyde has recently approved a strategic business project that concerns the processes of student feedback collection and how the work of student representatives can be supported though a designated application in a web-based or mobile format. The project is formally regarded as *StrathReps.* As part of the initial data gathering stage of a project, an evaluation of current procedures is carried out, with the data later being comparatively evaluated to evaluation data at the end of the project (Ward et al., 1996). For the *StrathReps* project, one approach for feedback collection presently is through a web-based ILE (interactive learning environment), called MyPlace, where students can submit feedback for their course or programme through a text-box, which allows relatively short-form textual representations, similar to a comment box on Facebook. This feedback can then be viewed by class/course representatives and lecturers but is also stored in a system database. The analysis of this textual data, using the prototype system developed is beneficial for a number of reasons, which will be detailed in the following paragraphs.

Through extracting the sentiment polarity of submitted feedback through the ILE, an overall tone of student feedback can be established, which can serve as an indicator of university, faculty or course performance (depending on the granularity of feedback). Furthermore, through identifying the polarity of historic feedback submissions, relevant measures can be taken to ensure the mental health wellbeing of class representatives and potentially lecturers that will use the system that will be designed. A predominantly negative feedback might require text manipulation prior to reporting to representatives or lecturers to avoid burnout, anxiety or depressive thoughts on any sort, arising from the submitted texts.

Through modelling topics, key areas that require improvement or recognition can be identified, which can be used as means of triangulation of feedback submitted directly to lecturers or directly through representatives (i.e. as opposed to through the ILE). Implementing an automatic, scalable solution to the identification of areas that require attention also enables data visualisation cross-faculty. This can lead to a reduction of response times on pressing issues, based on topic data visualisation, as well as potentially taking a proactive approach to quality-checking areas that are identified as problematic in two or more faculties. Collectively, this can result in an improved relationship between the university and students, which is assumed on the basis of feedback recognition and implementation.

Finally, at a project meeting that I attended it was brought up that there is a potential for a chat-bot type of system to be created as part of the project, to ease the feedback reporting process. To do so, a potential means of query classification was to prompt students to choose the type of feedback they would like to leave. Although there are many classification types, there are two that are common in academic literature: (1) structured and unstructured; and (2) positive, negative and intrinsic (see Vallerand and Reid, 1988; Shanab et al., 1981), yet other types exist e.g. corrective (Bitchener, 2008; Lyster and Ranta, 1997), behavioural and emotional (Damian et al., 2015; Jacobs et al., 1974) and so on. If text classification is performed using sample training data from the relevant feedback type, and similarity score as an evaluation metric, three things can be achieved: (1) an overview of the most common feedback types used by students can be extracted, which will subsequently lead to (2) a better understanding of the system requirements of the querying system for categoric display in the application, but also (3) a better understanding of the wording and linguistic expression of students will be achieved, which can help with the design of a more conversational-sounding chatbot, if one was to be created as part of the project.

Overall, from a holistic perspective the analysis for the *StrathReps* project requires a time- and cost-efficient solution for automation of the process of student-generated short-form text that can classify sentiment polarity, extract topics coherently and classify text in pre-defined categories, which completely matches the profile of the developed system as part of this research. Moreover,

from a marketing standpoint, the University can benefit from implementing automation, machine learning and learning analytics as part of a strategic profile as it is widely recognised in academic research that using available data to extract insight can be instrumental in strategic planning (Dziuban et al., 2012). The execution of this small-scale industrial research has been approved by the *StrathReps* project manager.

5.5.2 Application in Social media analytics for Hospitality and Tourism Industry

Another potential implementation of the system is in the area of hospitality and tourism, as discussed with company executives from MyCustomerLens. This industry arguably has a shorter customer journey from awareness to purchase, however if data is extracted from a combination of sources, such as social media, travel review websites, and company internal surveys and emails, analytics can demonstrate not only the topic-sentiment polarity, but also the consumers behavioural pattern across the customer journey in respect of digital outlet use. To elaborate, if access to the above listed data sources is available for marketing analytics organisations such as MyCustomerLens, an analytics dashboard can be built that demonstrates the user journey across digital outlets, showing for example if users that had a negative stay (sentiment polarity), which was caused by a long check in process (topic) would generally use twitter for writing a complaint (advocacy class; from the customer journey model) or write directly to the organisation. Such insight can be used the optimise the operational allocation of resources and improve business responsiveness.

5.5.3 Academic Research Experimental Opportunities for System Enhancement

As demonstrated by Section 3.4, in the Methodology chapter and Section 4.4 of the Analysis of results chapter, where the limitations of research and text classification results were discussed, respectively, one of the primary limitations of evidencing experimental results from supervised text classification was the lack of labels on the dataset, which obstructed testing and model evaluation. It is considered that this problem can be solved by the construction of an automated data extraction system that lacks the limitations of *Netlytic* (see Appendix D), and automatically assigns labels to extracted data on the basis of linguistic characteristics. Considering that these challenges are solved, a further route of improving the proposed system is the creation of digital identity patterns for text authors of various customer journey groups, informed by additional social media data points, such as the text author's following (i.e. number of followers), number of retweets and location information. Such data can be used to identify and prioritise responses, i.e. as means of PR (Personal Relations) strategy for damage control in highly influential cases and cases that can potentially impact the financial performance of the organisation negatively. Further

research opportunities were discussed throughout the text, and specifically in Section 3.3.3.2, in the Methodology Chapter.

5.6 **Personal Reflection Statement**

Experiential learning translates to learning from experience and personal practice, requiring critical reflection as means of encouraging the development and embedding of new skills and ways of thinking (Lewis and Williams, 1994). Therefore, a short reflection of the research process will be given. From an application standpoint, I consider this research being a success, yet I acknowledge the limitations it has in terms of scalability that stem from the tested algorithms being limited to the data at hand and the scope of this research. As a result of the research process I have gained an appreciation of the challenges, faced by industry professionals, trying to adapt machine learning and deep learning models to their unique business problems, which affirmed my passion for marketing process automation. This extension of my skill-set has, as observed by Schafersman (1991), resulted in a greater degree of self-motivation and critical thought, which I have demonstrated in both an academic and professional context, the latter being the application of the developed system in the University of Strathclyde's *StrathReps* project, from which I am currently an acting project assistant to.

5.7 **Conclusion**

This chapter presented conclusions in association with the research questions that were posed in earlier chapters and recommended action points for stakeholders. The future research section presented exiting opportunities, all of which offer benefits to existing business problems and some directly monetizable. Finally, please refer to the annex documentation, which is listed as Appendix K, where the documentation of supporting Python code for all associated experimental procedures is attached.

REFERENCES

Abid, F., Alam, M., Yasir, M. and Li, C., 2019. Sentiment analysis through recurrent variants latterly on convolutional neural network of Twitter. *Future Generation Computer Systems*, *95*, pp.292-308.

Abulaish, M. and Fazil, M., 2018. Modeling Topic Evolution in Twitter: An Embedding-Based Approach. *IEEE Access*, *6*, pp.64847-64857.

Aggarwal, C.C. and Zhai, C. eds., 2012. *Mining text data*. Springer Science & Business Media.

Ahuja, R., Chug, A., Kohli, S., Gupta, S. and Ahuja, P., 2019. The Impact of Features Extraction on the Sentiment Analysis. *Procedia Computer Science*, *152*, pp.341-348.

Ali, F., Kwak, D., Khan, P., El-Sappagh, S., Ali, A., Ullah, S., Kim, K.H. and Kwak, K.S., 2019. Transportation sentiment analysis using word embedding and ontology-based topic modeling. *Knowledge-Based Systems*, *In Press,* pp. 1-16.

Almeida, A.M., Cerri, R., Paraiso, E.C., Mantovani, R.G. and Junior, S.B., 2018. Applying multi-label techniques in emotion identification of short texts. *Neurocomputing*, *320*, pp.35-46.

Amaravadi, C.S., Samaddar, S. and Dutta, S., 1995. Intelligent marketing information systems: computerized intelligence for marketing decision making. *Marketing Intelligence & Planning*, *13*(2), pp.4-13.

Ampofo, L., Collister, S., O'Loughlin, B., Chadwick, A., Halfpenny, P.J. and Procter, P.J., 2015. Text mining and social media: When quantitative meets qualitative and software meets people. *Innovations in digital research methods. Thousand Oaks, CA: SAGE Publications Inc*, pp.161-92.

Appel, O., Chiclana, F., Carter, J. and Fujita, H., 2016. A hybrid approach to the sentiment analysis problem at the sentence level. *Knowledge-Based Systems*, *108*, pp.110-124.

Araque, O., Corcuera-Platas, I., Sanchez-Rada, J.F. and Iglesias, C.A., 2017. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, *77*, pp.236-246.

Asuncion, A., Welling, M., Smyth, P. and Teh, Y.W., 2009, June. On smoothing and inference for topic models. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence* (pp. 27-34). AUAI Press.

Baccianella, S., Esuli, A. and Sebastiani, F., 2010, May. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th Conference on Language Resources and Evaluation LREC* (Vol. 10, No. 2010, pp. 2200-2204).

Banea, C., Mihalcea, R. and Wiebe, J., 2008, May. A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources. In *Proceedings of the 6th International Conference on Language Resources and Evaluation LREC* (Vol. 8, pp. 2-764).

Bastani, K., Namavari, H. and Shaffer, J., 2019. Latent Dirichlet Allocation (LDA) for Topic Modeling of the CFPB Consumer Complaints. *Expert Systems with Applications*, *127,* pp. 256-271.

Baziotis, C., Pelekis, N. and Doulkeridis, C., 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 747-754).

Bell, E., Bryman, A. and Harley, B., 2018. *Business research methods*. Oxford: Oxford university press.

Bergamaschi, S. and Po, L., 2014, April. Comparing LDA and LSA topic models for content-based movie recommendation systems. In *International Conference on Web Information Systems and Technologies* (pp. 247-263). Springer, Cham.

Bernstein, R.J., 2011. *Beyond objectivism and relativism: Science, hermeneutics, and praxis*. University of Pennsylvania Press.

Berry, M.J. and Linoff, G.S., 2004. *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons: New York, USA.

Bhadane, C., Dalal, H. and Doshi, H., 2015. Sentiment analysis: Measuring opinions. *Procedia Computer Science*, *45*, pp.808-814.

Bird, S., Klein, E. and Loper, E., 2009. *Natural Language Processing with Python.* O'Reilly Media: California, USA.

Bitchener, J., 2008. Evidence in support of written corrective feedback. *Journal of second language writing*, *17*(2), pp.102-118.

Blei, D.M. and Lafferty, J.D., 2007. A correlated topic model of science. *The Annals of Applied Statistics*, *1*(1), pp.17-35.

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), pp.993-1022.

Bloomberg, L.D. and Volpe, M., 2018. *Completing your qualitative dissertation: A road map from beginning to end*. London: Sage Publications.

Bonzanini, M., 2016. *Mastering social media mining with Python*. Packt Publishing Ltd.

Booth, A., Sutton, A. and Papaioannou, D., 2016. *Systematic approaches to a successful literature review*, 2nd Edition. London, UK: Sage Publications.

Bravo-Marquez, F., Mendoza, M. and Poblete, B., 2014. Meta-level sentiment models for big social data analysis. *Knowledge-Based Systems*, *69*, pp.86-99.

Brownlee, J., 2017. *How to Plan and Run Machine Learning Experiments Systematically.* Machine Learning Mastery [blog] Available on: https://machinelearningmastery.com/plan-run-machine-learning-experiments-systematically/ [Accessed on 11.03.2019]

Bucak, S.S., Jin, R. and Jain, A.K., 2013. Multiple kernel learning for visual object recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(7), pp.1354-1369.

Buchanan, D. and Bryman, A. eds., 2009. *The Sage handbook of organizational research methods*. Thousand Oaks, CA: Sage Publications Ltd.

CABS, 2019. *AJG 2018*, Online. Available from: https://charteredabs.org/academic-journal-guide-2018/ [Accessed 16.06.2019]

Camacho-Collados, J. and Pilehvar, M.T., 2017. On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis. *arXiv preprint arXiv:1707.01780*.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L. and Blei, D.M., 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*(pp. 288-296).

Chaturvedi, I., Cambria, E., Welsch, R.E. and Herrera, F., 2018. Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion*, *44*, pp.65-77.

Chaturvedi, I., Ong, Y.S., Tsang, I.W., Welsch, R.E. and Cambria, E., 2016. Learning word dependencies in text by means of a deep recurrent belief network. *Knowledge-Based Systems*, *108*, pp.144-154.

Chawla, N.V., Japkowicz, N. and Kotcz, A., 2004. Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, *6*(1), pp.1-6.

Chen, C.H., Lee, W.P. and Huang, J.Y., 2018. Tracking and recognizing emotions in short text messages from online chatting services. *Information Processing & Management*, *54*(6), pp.1325-1344.

Chen, F., Ji, R., Su, J., Cao, D. and Gao, Y., 2017. Predicting microblog sentiments via weakly supervised multimodal deep learning. *IEEE Transactions on Multimedia*, *20*(4), pp.997-1007.

Chen, H., Chiang, R.H. and Storey, V.C., 2012. Business intelligence and analytics: From big data to big impact. *MIS quarterly*, *36*(4).

Chen, M., Jin, X. and Shen, D., 2011, June. Short text classification improved by learning multi-granularity topics. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

Chen, P., Zhang, N.L., Liu, T., Poon, L.K., Chen, Z. and Khawar, F., 2017. Latent tree models for hierarchical topic detection. *Artificial Intelligence*, *250*, pp.105-124.

Chen, Q., Guo, X. and Bai, H., 2017. Semantic-based topic detection using Markov decision processes. *Neurocomputing*, *242*, pp.40-50.

Chen, Y., Zhang, H., Liu, R., Ye, Z. and Lin, J., 2019. Experimental explorations on short text topic mining between LDA and NMF based Schemes. *Knowledge-Based Systems*, *163*, pp.1-13.

Chiu, C.Y., Gelfand, M.J., Yamagishi, T., Shteynberg, G. and Wan, C., 2010. Intersubjective culture: The role of intersubjective perceptions in cross-cultural research. *Perspectives on Psychological Science*, *5*(4), pp.482-493.

Chollet, F., 2018. *Deep learning with Python.* Manning: Shelter Island, NY.

Chowdhury, A., 2019. Design and Development of a Stencil for Mobile User Interface (UI) Design. In *Research into Design for a Connected World* (pp. 629-639). Springer, Singapore.

Chuang, J., Manning, C.D. and Heer, J., 2012, May. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the international working conference on advanced visual interfaces* (pp. 74-77). ACM.

Cigarrán, J., Castellanos, Á. and García-Serrano, A., 2016. A step forward for Topic Detection in Twitter: An FCA-based approach. *Expert Systems with Applications*, *57*, pp.21-36.

Clarivate Analytics, 2019. *InCites Journal Citation Reports*, Online. Available from: jcr.clarivate.com/JCRLandingPageAction.action?Init=Yes&SrcApp=IC2LS&SID=J4-rUPQW7H2KbOTXo7Kbwx66BA0ScWilKIaMPF-H3SgPCcRKv710YZayRLBE8NJors5fOF8CaPEiZNp27HcJMrRRzwptyuRc40kCVKnXE-qBgNuLRjcgZrPm66fhjx2Fmwx3Dx3D-h9tQNJ9Nv4eh45yLvkdX3gx3Dx3D#Clar [Accessed 16.06.2019]

Clark, E. and Araki, K., 2011. Text normalization in social media: progress, problems and applications for a pre-processing system of casual English. *Procedia-Social and Behavioral Sciences*, *27*, pp.2-11.

Clarke, K. and Belk, R.W., 1979. The effects of product involvement and task definition on anticipated consumer effort. *ACR North American Advances*.

Clavel, C. and Callejas, Z., 2015. Sentiment analysis: from opinion mining to human-agent interaction. *IEEE Transactions on affective computing*, *7*(1), pp.74-93.

Collins, K., 2010. Advanced sampling designs in mixed research: Current practices and emerging trends in the social and behavioral sciences. *Sage handbook of mixed methods in social and behavioral research*, *2*, pp.353-377.

Converse, J.M. and Presser, S., 1986. *Survey questions: Handcrafting the standardized questionnaire* (No. 63). Sage.

Cooper, D. R. and Schindler, P.S., 2014. *Business Research Methods*. 12th Edition. Irwin: McGraw-Hill

Crannell, W.C., Clark, E., Jones, C., James, T.A. and Moore, J., 2016. A pattern-matched Twitter analysis of US cancer-patient sentiments. *journal of surgical research*, *206*(2), pp.536-542.

Curiskis, S.A., Drake, B., Osborn, T.R. and Kennedy, P.J., 2019. An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. *Information Processing & Management*, *In Press.*

da Silva, N.F.F., Coletta, L.F., Hruschka, E.R. and Hruschka Jr, E.R., 2016. Using unsupervised information to improve semi-supervised tweet sentiment classification. *Information Sciences*, *355*, pp.348-365.

Damian, I., Tan, C.S.S., Baur, T., Schöning, J., Luyten, K. and André, E., 2015, April. Augmenting social interactions: Realtime behavioural feedback using social signal

processing techniques. In *Proceedings of the 33rd annual ACM conference on Human factors in computing systems* (pp. 565-574). ACM.

Dang, Q., Gao, F. and Zhou, Y., 2016. Early detection method for emerging topics based on dynamic bayesian networks in micro-blogging networks. *Expert Systems with Applications*, *57*, pp.285-295.

Danneman, N. and Heimann, R., 2014. *Social media mining with R*. Packt Publishing Ltd.

Davidson, D., 1996. Subjective, intersubjective, objective. In *Philosophy*. 3rd Volume. Oxford University Press: Bristol: Thoemmes. pp. 555-558

Deloitte, 2019. *2019 Insurance Industry Outlook*, Deloitte, Online. Available at: https://www2.deloitte.com/us/en/pages/financial-services/articles/insurance-industry-outlook.html [Accessed 18.08.2019]

Diamantini, C., Mircoli, A., Potena, D. and Storti, E., 2019. Social information discovery enhanced by sentiment analysis techniques. *Future Generation Computer Systems*, *95*, pp.816-828.

Dillman, D.A. and Bowker, D.K., 2001. The web questionnaire challenge to survey methodologists. *Online social sciences*, pp.53-71.

Dodig-Crnkovic, G., 2002, April. Scientific methods in computer science. In *Proceedings of the Conference for the Promotion of Research in IT at New Universities and at University Colleges in Sweden, Skövde, Suecia* (pp. 126-130).

Dohaiha, H.H., Prasad, P.W.C., Maag, A. and Alsadoon, A., 2018. Deep learning for aspect-based sentiment analysis: a comparative review. *Expert Systems with Applications*.

Dong, L.Y., Ji, S.J., Zhang, C.J., Zhang, Q., Chiu, D.W., Qiu, L.Q. and Li, D., 2018. An unsupervised topic-sentiment joint probabilistic model for detecting deceptive reviews. *Expert Systems with Applications*, *114*, pp.210-223.

Dos Santos, C. and Gatti, M., 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 69-78).

Dos Santos, F.L. and Ladeira, M., 2014, October. The role of text pre-processing in opinion mining on a social media language dataset. In *2014 Brazilian Conference on Intelligent Systems* (pp. 50-54). IEEE.

Dziuban, C., Moskal, P., Cavanagh, T. and Watts, A., 2012. Analytics that Inform the University: Using Data You Already Have. *Journal of Asynchronous Learning Networks*, *16*(3), pp.21-38.

Echeverry-Correa, J.D., Ferreiros-López, J., Coucheiro-Limeres, A., Córdoba, R. and Montero, J.M., 2015. Topic identification techniques applied to dynamic language model adaptation for automatic speech recognition. *Expert Systems with Applications*, *42*(1), pp.101-112.

Einwiller, S.A. and Steilen, S., 2015. Handling complaints on social network sites–An analysis of complaints and complaint responses on Facebook and Twitter pages of large US companies. *Public Relations Review*, *41*(2), pp.195-204.

El-Diraby, T., Shalaby, A. and Hosseini, M., 2019. Linking social, semantic and sentiment analyses to support modeling transit customers' satisfaction: Towards formal study of opinion dynamics. *Sustainable Cities and Society*, *49*, p.101578.

Fan, W. and Gordon, M.D., 2014. The power of social media analytics. *Communications of the ACM*, *57*(6), pp.74-81.

Farhadloo, M., Patterson, R.A. and Rolland, E., 2016. Modeling customer satisfaction from unstructured data using a Bayesian approach. *Decision Support Systems*, *90*, pp.1-11.

Feeney, A. and Heit, E. eds., 2007. *Inductive reasoning: Experimental, developmental, and computational approaches*. Cambridge University Press.

Fernández-Gavilanes, M., Álvarez-López, T., Juncal-Martínez, J., Costa-Montenegro, E. and González-Castaño, F.J., 2016. Unsupervised method for sentiment analysis in online texts. *Expert Systems with Applications*, *58*, pp.57-75.

Fernandez-Lozano, C., Gestal, M., Munteanu, C.R., Dorado, J. and Pazos, A., 2016. A methodology for the design of experiments in computational intelligence with multiple regression models. *PeerJ*, *4*, p.e2721.

Fink, A., 2019. *Conducting research literature reviews: from the Internet to paper*, 2nd Edition. Thousand Oaks, California: Sage Publications.

Flick, U., 2014. *An introduction to qualitative research*, 5th Edition. Sage: London, UK.

Følstad, A. and Kvale, K., 2018. Customer journeys: a systematic literature review. *Journal of Service Theory and Practice*, *28*(2), pp.196-227.

Fu, X., Sun, X., Wu, H., Cui, L. and Huang, J.Z., 2018. Weakly supervised topic sentiment joint model with word embeddings. *Knowledge-Based Systems*, *147*, pp.43-54.

Fu, X., Wei, Y., Xu, F., Wang, T., Lu, Y., Li, J. and Huang, J.Z., 2019. Semi-supervised Aspect-level Sentiment Classification Model based on Variational Autoencoder. *Knowledge-Based Systems*, *171*, pp.81-92.

Fu, X., Yang, K., Huang, J.Z. and Cui, L., 2015. Dynamic non-parametric joint sentiment topic mixture model. *Knowledge-Based Systems*, *82*, pp.102-114.

García, S., Fernández, A., Luengo, J. and Herrera, F., 2010. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*, *180*(10), pp.2044-2064.

GDPR, 2018.*General Data Protection Regulation,* Online. Available: https://gdpr-info.eu [Accessed 30.07.2019]

Géron, A., 2017. *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. O'Reilly Media: California, USA

Goodfellow, I., Bengio, Y. and Courville, A., 2016. *Deep learning*. MIT press.

Goodson, L. and Phillimore, J. eds., 2004. *Qualitative research in tourism: Ontologies, epistemologies and methodologies*. London: Routledge.

Gruzd, A., 2016. Netlytic: Software for automated text and social network analysis. *Computer software,* Available from: http://Netlytic. org [accessed 28.07.2019]

Gruzd, A., Mai, P. and Kampen, A., 2016. A how-to for using Netlytic to collect and analyze social media data: A case study of the use of Twitter during the 2014 Euromaidan Revolution in Ukraine. *The SAGE handbook of social media research methods*, pp.513-529.

Gupta, B., Negi, M., Vishwakarma, K., Rawat, G. and Badhani, P., 2017. Study of Twitter sentiment analysis using machine learning algorithms on Python. *International Journal of Computer Applications*, *165*(9), pp.0975-8887.

Hajjem, M. and Latiri, C., 2017. Combining IR and LDA topic modeling for filtering microblogs. *Procedia Computer Science*, *112*, pp.761-770.

Hardeniya, N., Perkins, J., Chopra, D., Joshi, N. and Mathur, I., 2016. *Natural Language Processing: Python and NLTK*. Packt Publishing Ltd.

Hart, C., 2018. *Doing a literature review: Releasing the research imagination*, 2nd Edition. London, UK: Sage Publications.

Howells, K. and Ertugan, A., 2017. Applying fuzzy logic for sentiment analysis of social media network data in marketing. *Procedia computer science*, *120*, pp.664-670.

Huang, F., Zhang, S., Zhang, J. and Yu, G., 2017. Multimodal learning for topic sentiment analysis in microblogging. *Neurocomputing*, *253*, pp.144-153.

Ibrahim, N.F. and Wang, X., 2019. A text analytics approach for online retailing service improvement: Evidence from Twitter. *Decision Support Systems*, *121*, pp.37-50.

IEEE Computer Society, 1998. 830-1998 - IEEE Recommended Practice for Software Requirements Specifications. *IEEE*

Ittoo, A., Nguyen, L.M. and van den Bosch, A., 2016. Text analytics in industry: Challenges, desiderata and trends. *Computers in Industry*, *78*, pp.96-107.

Jacobs, A., Jacobs, M., Cavior, N. and Burke, J., 1974. Anonymous feedback: Credibility and desirability of structured emotional and behavioral feedback delivered in groups. *Journal of Counseling Psychology*, *21*(2), p.106.

Japkowicz, N. and Shah, M., 2011. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press.

Jesson, J., Matheson, L. and Lacey, F.M., 2011. *Doing your literature review: Traditional and systematic techniques*. London, UK: Sage Publications.

Jiang, B., Li, Z., Chen, H. and Cohn, A.G., 2018. Latent Topic Text Representation Learning on Statistical Manifolds. *IEEE transactions on neural networks and learning systems*, *29*(11), pp. 5643-5654.

Johnson, J. M. and Rowlands, T., 2012. The Interpersonal Dynamics of In-Depth Interviewing. In J. Gubrium, J. Holstein, A. Marvasti and K. McKinney, ed. *The SAGE Handbook of Interview Research,* 2nd ed. United States of America: SAGE

Johnson, R.B. and Onwuegbuzie, A.J., 2004. Mixed methods research: A research paradigm whose time has come. *Educational researcher*, *33*(7), pp.14-26.

Joshi, P. 2018. *Text Mining 101: A Stepwise Introduction to Topic Modeling using Latent Semantic Analysis (using Python).* Analytics Vidhya [blog] Available from: https://www.analyticsvidhya.com/blog/2018/10/stepwise-guide-topic-modeling-latent-semantic-analysis/ [Accessed 29.07.2019]

Kayser, V. and Bierwisch, A., 2016. Using Twitter for foresight: An opportunity?. *Futures*, *84*, pp.50-63.

Kemp, S., 2019. Digital 2019: *Global Internet Use Accelerates.* We Are Social: Hootsuite Special Report, Online. Available at: https://wearesocial.com/blog/2019/01/digital-2019-global-internet-use-accelerates [Accessed 15.07.2019]

Khan, F.H., Bashir, S. and Qamar, U., 2014. TOM: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems*, *57*, pp.245-257.

Kim, J. and Sung, Y., 2009. Dimensions of purchase-decision involvement: Affective and cognitive involvement in product and brand. *Journal of Brand Management*, *16*(8), pp.504-519.

King, D.B., O'Rourke, N. and DeLongis, A., 2014. Social media recruitment and online data collection: A beginner's guide and best practices for accessing low-prevalence and hard-to-reach populations. *Canadian Psychology/Psychologie canadienne*, *55*(4), p.240.

Kintsch, W., McNamara, D.S., Dennis, S. and Landauer, T.K., 2007. LSA and meaning: In theory and application. *Handbook of latent semantic analysis*. Lawrence Erlbaum, Mahwah, NJ, USA, pp.467-479.

Korfiatis, N., Stamolampros, P., Kourouthanassis, P. and Sagiadinos, V., 2019. Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews. *Expert Systems with Applications*, *116*, pp.472-486.

Lee, T.Y., Smith, A., Seppi, K., Elmqvist, N., Boyd-Graber, J. and Findlater, L., 2017. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies*, *105*, pp.28-42.

Lemon, K.N. and Verhoef, P.C., 2016. Understanding customer experience throughout the customer journey. *Journal of marketing*, *80*(6), pp.69-96.

Lewis, J. and Ritchie, J., 2003. Generalising from qualitative research. In *Qualitative research practice – a guide for social science students and researchers*. Ed. J. Ritchie and J. Lewis, pp. 263-286. London: Sage Publications.

Lewis, L.H. and Williams, C.J., 1994. Experiential learning: Past and present. *New directions for adult and continuing education*, *1994*(62), pp.5-16.

Li, J., Wu, N. and Feng, Z., 2018. Model-based non-gaussian interest topic distribution for user retweeting in social networks. *Neurocomputing*, *278*, pp.87-98.

Li, X., Wu, C. and Mai, F., 2019. The effect of online reviews on product sales: A joint sentiment-topic analysis. *Information & Management*, *56*(2), pp.172-184.

Li, X., Zhang, A., Li, C., Ouyang, J. and Cai, Y., 2018. Exploring coherent topics by topic modeling with term weighting. *Information Processing & Management*, *54*(6), pp.1345-1358.

Li, Y., Guo, H., Zhang, Q., Gu, M. and Yang, J., 2018. Imbalanced text sentiment classification using universal and domain-specific knowledge. *Knowledge-Based Systems*, *160*, pp.1-15.

Liang, W., Xie, H., Rao, Y., Lau, R.Y. and Wang, F.L., 2018. Universal affective model for Readers' emotion classification over short texts. *Expert Systems with Applications*, *114*, pp.322-333.

Lin, L.Y. and Chen, C.S., 2006. The influence of the country-of-origin image, product knowledge and product involvement on consumer purchase decisions: an empirical study of insurance and catering services in Taiwan. *Journal of consumer Marketing*, *23*(5), pp.248-265.

Linoff, G.S. and Berry, M.J., 2011. *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons: Canada. 2nd Edition.

Littman, J., 2017. *Where to get Twitter data for academic research.* Social Feed Manager [online] Available on: https://gwu-libraries.github.io/sfm-ui/posts/2017-09-14-twitter-data [Accessed 11.03.2019]

Liu, B. and Zhang, L., 2012. A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415-463). Springer US.

Liu, Y., Wang, J. and Jiang, Y., 2016. PT-LDA: A latent variable model to predict personality traits of social network users. *Neurocomputing*, *210*, pp.155-163.

Lo, S.L., Chiong, R. and Cornforth, D., 2017. An unsupervised multilingual approach for online social media topic identification. *Expert Systems with Applications*, *81*, pp.282-298.

Lozano, M.G., Schreiber, J. and Brynielsson, J., 2017. Tracking geographical locations using a geo-aware topic model for analyzing social media data. *Decision Support Systems*, *99*, pp.18-29.

Lyster, R. and Ranta, L., 1997. Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in second language acquisition*, *19*(1), pp.37-66.

Madhoushi, Z., Hamdan, A.R. and Zainudin, S., 2015, July. Sentiment analysis techniques in recent works. In *2015 Science and Information Conference (SAI)* (pp. 288-291). IEEE.

Majumder, N., Hazarika, D., Gelbukh, A., Cambria, E. and Poria, S., 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-Based Systems*, *161*, pp.124-133.

Mao, Y. and Lebanon, G., 2007. Isotonic conditional random fields and local sentiment flow. In *Advances in neural information processing systems* (pp. 961-968).

Mason, J., 2002. *Qualitative researching*. 2nd Ed. London: Sage Publications.

Mei, Q., Ling, X., Wondra, M., Su, H. and Zhai, C., 2007, May. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web* (pp. 171-180). ACM.

Mimno, D., Wallach, H.M., Talley, E., Leenders, M. and McCallum, A., 2011, July. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 262-272). Association for Computational Linguistics.

Mittal, B., 1989. Measuring purchase-decision involvement. *Psychology & Marketing*, *6*(2), pp.147-162.

Moher, D., Liberati, A., Tetzlaff, J. and Altman, D.G., 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine*, *151*(4), pp.264-269.

Nadkarni, P.M., Ohno-Machado, L. and Chapman, W.W., 2011. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5), pp.544-551.

Nakatani, K. and Chuang, T.T., 2011. A web analytics tool selection method: an analytical hierarchy process approach. *Internet Research*, *21*(2), pp.171-186.

Nassirtoussi, A.K., Aghabozorgi, S., Wah, T.Y. and Ngo, D.C.L., 2014. Text mining for market prediction: A systematic review. *Expert Systems with Applications*, *41*(16), pp.7653-7670.

Nielsen, M.A., 2015. *Neural networks and deep learning* (Vol. 25). Determination press: USA

Norris, C., 2005. *Epistemology: Key concepts in philosophy*. London: A&C Black.

Okoli, C. and Schabram, K., 2010. A guide to conducting a systematic literature review of information systems research. *Sprouts: Working Papers on Information Systems, 10*(26), pp. 1-50.

Omand, D., Bartlett, J. and Miller, C., 2012. Introducing social media intelligence (SOCMINT). *Intelligence and National Security*, *27*(6), pp.801-823.

Onan, A., Korukoglu, S. and Bulut, H., 2016. LDA-based Topic Modelling in Text Sentiment Classification: An Empirical Analysis. *Int. J. Comput. Linguistics Appl.*, *7*(1), pp.101-119.

Oza, K.S. and Naik, P.G., 2016. Prediction of online lectures popularity: a text mining approach. *Procedia Computer Science*, *92*, pp.468-474.

Öztürk, N. and Ayvaz, S., 2018. Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis. *Telematics and Informatics*, *35*(1), pp.136-147.

Pang, B. and Lee, L., 2004, July. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics* (p. 271). Association for Computational Linguistics.

Pang, B., and Lee, L. 2008. "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information* Retrieval (2:1-2), pp. 1-135.

Park, D.H., Lee, J. and Han, I., 2007. The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement. *International journal of electronic commerce*, *11*(4), pp.125-148.

Patton, M.Q., 2002. *Qualitative research and evaluation methods*. 3rd ed. Thousand Oaks: Sage.

Pennington, J., Socher, R. and Manning, C., 2014, October. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Pröllochs, N. and Feuerriegel, S., 2018. Business analytics for strategic management: Identifying and assessing corporate challenges via topic modeling. *Information & Management*.

Qian, S., Zhang, T., Xu, C. and Shao, J., 2015. Multi-modal event topic model for social event analysis. *IEEE transactions on multimedia*, *18*(2), pp.233-246.

Qiu, L., Lei, Q. and Zhang, Z., 2018. Advanced sentiment classification of tibetan microblogs on smart campuses based on multi-feature fusion. *IEEE Access*, *6*, pp.17896-17904.

Quinlan, C., Babin, B., Carr, J. and Griffin, M., 2019. *Business research methods*. South Western Cengage.

Rao, Y., Xie, H., Li, J., Jin, F., Wang, F.L. and Li, Q., 2016. Social emotion classification of short text via topic-level maximum entropy model. *Information & Management*, *53*(8), pp.978-986.

Rawson, A., Duncan, E. and Jones, C., 2013. The truth about customer experience. *Harvard Business Review*, *91*(9), pp.90-98.

Ren, Y., Wang, R. and Ji, D., 2016. A topic-enhanced word embedding for Twitter sentiment classification. *Information Sciences*, *369*, pp.188-198.

Rife, S.C., Cate, K.L., Kosinski, M. and Stillwell, D., 2016. Participant recruitment and data collection through Facebook: The role of personality factors. *International Journal of Social Research Methodology*, *19*(1), pp.69-83.

Rish, I., 2001, August. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, No. 22, pp. 41-46).

Ritter, A., Cherry, C. and Dolan, W.B., 2011. Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 583-593). July. Association for Computational Linguistics.

Rosa, R.L., Schwartz, G.M., Ruggiero, W.V. and Rodriguez, D.Z., 2018. A Knowledge-Based Recommendation System that includes Sentiment Analysis and Deep Learning. *IEEE Transactions on Industrial Informatics*, *15*(4), pp. 2124-2135.

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A. and Stoyanov, V., 2015, June. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*(pp. 451-463).

Rossmann, A. and Stei, G., 2015. User Engagement in Corporate Facebook Communities. In: Zimmermann, A. & Rossmann, A. (Ed.) *Digital Enterprise Computing* (Dec 2015). Bonn: Gesellschaft für Informatik e.V., pp. 51-62.

Rousseau, D.M., Manning, J. and Denyer, D., 2008. Evidence in management and organizational science: assembling the field's full weight of scientific knowledge through syntheses. *The academy of management annals*, *2*(1), pp.475-515.

Russell, S.J. and Norvig, P., 2016. *Artificial intelligence: a modern approach*. Pearson Education Limited: Malaysia.

Rybalko, S. and Seltzer, T., 2010. Dialogic communication in 140 characters or less: How Fortune 500 companies engage stakeholders using Twitter. *Public relations review*, *36*(4), pp.336-341.

Salloum, S.A., Al-Emran, M., Monem, A.A. and Shaalan, K., 2017. A survey of text mining in social media: facebook and twitter perspectives. *Adv. Sci. Technol. Eng. Syst. J*, *2*(1), pp.127-133.

Sangiorgi, D., 2011. *Transformative services and transformation design,* Gower.

Sapountzi, A. and Psannis, K.E., 2018. Social networking data analysis tools & challenges. *Future Generation Computer Systems*, *86*, pp.893-913.

Sarantakos, S., 2012. *Social research*. Macmillan International Higher Education.

Sarkar, D., 2016. *Text Analytics with Python*. Apress.

Saunders, M.L., Lewis, P.P. and Thornhill, A., 2016. *Research methods for business students,* 7th edition. Pearson Education: Harlow

Sayyadi, H. and Raschid, L., 2013. A graph analytical approach for topic detection. *ACM Transactions on Internet Technology (TOIT)*, *13*(2), p.4.

Schafersman, S.D., 1991. An introduction to critical thinking.

Schwandt, T.A, 2001. *Dictionary of qualitative inquiry.* 2nd Edition. Thousand Oaks: Sage.

Shanab, M.E., Peterson, D., Dargahi, S. and Deroian, P., 1981. The effects of positive and negative verbal feedback on the intrinsic motivation of male and female subjects. *The Journal of Social Psychology*, *115*(2), pp.195-205.

Sintsova, V. and Pu, P., 2016. Dystemo: Distant supervision method for multi-category emotion recognition in tweets. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *8*(1), p.13.

Soldatova, L.N. and King, R.D., 2006. An ontology of scientific experiments. *Journal of the Royal Society Interface*, *3*(11), pp.795-803.

Song, G., Ye, Y., Du, X., Huang, X. and Bie, S., 2014. Short text classification: A survey. *Journal of multimedia*, *9*(5), p.635.

Srinivasa-Desikan, B., 2018. *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing Ltd.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H. and Demirbas, M., 2010, July. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 841-842). ACM.

Srividhya, V. and Anitha, R., 2010. Evaluating preprocessing techniques in text categorization. International journal of computer science and application, 47(11), pp.49-51.

Stebbins, R.A., 2001. *Exploratory research in the social sciences* (Vol. 48). London: Sage.

Stevens, K., Kegelmeyer, P., Andrzejewski, D. and Buttler, D., 2012, July. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 952-961). Association for Computational Linguistics.

Stickdorn, M. and Schneider, J., 2010. *This is Service Design Thinking. Basics, Tools, Cases*. 5. BIS Publishers, Amsterdam.

Strapparava, C. and Valitutti, A., 2004, May. Wordnet affect: an affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation LREC* (Vol. 4, No. 1083-1086, p. 40).

Sun, F., Belatreche, A., Coleman, S., McGinnity, T.M. and Li, Y., 2014, March. Pre-processing online financial text for sentiment classification: A natural language processing approach. In *2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)* (pp. 122-129). IEEE.

Sun, X., Li, C. and Ren, F., 2016. Sentiment analysis for Chinese microblog based on deep neural networks with convolutional extension features. *Neurocomputing*, *210*, pp.227-236.

Symeonidis, S., Effrosynidis, D. and Arampatzis, A., 2018. A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications*, *110*, pp.298-310.

Tang, D., Qin, B. and Liu, T., 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1422-1432).

Tang, D., Zhang, Z., He, Y., Lin, C. and Zhou, D., 2019. Hidden topic–emotion transition model for multi-level social emotion detection. *Knowledge-Based Systems*, *164*, pp.426-435.

Taylor, J. and Pagliari, C., 2018. Mining social media data: How are research sponsors and researchers addressing the ethical challenges?. *Research Ethics*, *14*(2), pp.1-39.

Thanaki, J., 2017. *Python Natural Language Processing*. Packt Publishing Ltd.

Tichy, W.F., 1998. Should computer scientists experiment more?. *Computer*, *31*(5), pp.32-40.

Tongco, M.D.C., 2007. Purposive sampling as a tool for informant selection. *Ethnobotany Research and applications*, *5*, pp.147-158.

Uys, J.W., Du Preez, N.D. and Uys, E.W., 2008, July. Leveraging unstructured information using topic modelling. In *PICMET'08-2008 Portland International Conference on Management of Engineering & Technology* (pp. 955-961). IEEE.

Vallerand, R.J. and Reid, G., 1988. On the relative effects of positive and negative verbal feedback on males' and females' intrinsic motivation. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, *20*(3), p.239.

Vijayarani, S. and Janani, R., 2016. Text mining: open source tokenization tools-an analysis. *Advanced Computational Intelligence: An International Journal (ACII)*, *3*(1), pp.37-47.

Voorhees, C.M., Fombelle, P.W., Gregoire, Y., Bone, S., Gustafsson, A., Sousa, R. and Walkowiak, T., 2017. Service encounters, experiences and the customer journey: Defining the field and a call to expand our lens. *Journal of Business Research*, *79*, pp.269-280.

Vulić, I., De Smet, W., Tang, J. and Moens, M.F., 2015. Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing & Management*, *51*(1), pp.111-147.

Wallach, H.M., Murray, I., Salakhutdinov, R. and Mimno, D., 2009, June. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning* (pp. 1105-1112). ACM.

Wang, F.Y., Carley, K.M., Zeng, D. and Mao, W., 2007. Social computing: From social informatics to social intelligence. *IEEE Intelligent Systems*, 22(2).

Wang, G., Sun, J., Ma, J., Xu, K. and Gu, J., 2014. Sentiment classification: The contribution of ensemble learning. *Decision support systems*, *57*, pp.77-93.

Wang, H.C., Jhou, H.T. and Tsai, Y.S., 2018. Adapting topic map and social influence to the personalized hybrid recommender system. *Information Sciences*, *In Press.*

Wang, P., Xu, B., Xu, J., Tian, G., Liu, C.L. and Hao, H., 2016. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. *Neurocomputing*, *174*, pp.806-814.

Wang, P., Xu, J., Xu, B., Liu, C., Zhang, H., Wang, F. and Hao, H., 2015. Semantic clustering and convolutional neural network for short text categorization. In *Proceedings of the*

53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (Vol. 2, pp. 352-357).

Wang, W., Feng, Y. and Dai, W., 2018. Topic analysis of online reviews for two competitive products using latent Dirichlet allocation. *Electronic Commerce Research and Applications*, *29*, pp.142-156.

Ward, J., Taylor, P. and Bond, P., 1996. Evaluation and realisation of IS/IT benefits: an empirical study of current practice. *European Journal of Information Systems*, *4*(4), pp.214-225.

Wells, J.D., Valacich, J.S. and Hess, T.J., 2011. What signal are you sending? How website quality influences perceptions of product quality and purchase intentions. *MIS quarterly*, pp.373-396.

Xiao, Y., Fan, Z., Tan, C., Xu, Q., Zhu, W. and Cheng, F., 2019. Sense-Based Topic Word Embedding Model for Item Recommendation. *IEEE Access*, *7*, pp.44748-44760.

Xiaomei, Z., Jing, Y., Jianpei, Z. and Hongyu, H., 2018. Microblog sentiment analysis with weak dependency connections. *Knowledge-Based Systems*, *142*, pp.170-180.

Xiong, S., Wang, K., Ji, D. and Wang, B., 2018. A short text sentiment-topic model for product reviews. *Neurocomputing*, *297*, pp.94-102.

Xu, S., 2018. Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science*, *44*(1), pp.48-59.

Xu, Y., Yin, Y. and Yin, J., 2017. Tackling topic general words in topic modeling. *Engineering Applications of Artificial Intelligence*, *62*, pp.124-133.

Yan, X., Guo, J., Lan, Y. and Cheng, X., 2013, May. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 1445-1456). ACM.

Yan, Y., Yang, H. and Wang, H.M., 2017, July. Two simple and effective ensemble classifiers for Twitter sentiment analysis. In *2017 Computing Conference* (pp. 1386-1393). IEEE.

Yang, Y., Pierce, T. and Carbonell, J.G., 1998. A study on retrospective and on-line event detection. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp.28-36.

Yin, R., 2003. Case study research: design and methods. 3rd Ed. Thousand Oaks, CA: Sage Publications.

Yu, D., Xu, D., Wang, D. and Ni, Z., 2019. Hierarchical Topic Modeling of Twitter Data for Online Analytical Processing. *IEEE Access*, *7*, pp.12373-12385.

Yu, J. and Qiu, L., 2018. ULW-DMM: An effective topic modeling method for microblog short text. *IEEE Access*, *7*, pp.884-893.

Zhang, C., Wang, H., Cao, L., Wang, W. and Xu, F., 2016. A hybrid term–term relations analysis approach for topic detection. *Knowledge-Based Systems*, *93*, pp.109-120.

Zhang, H. and Zhong, G., 2016. Improving short text classification by learning vector representations of both words and hidden topics. *Knowledge-Based Systems*, *102*, pp.76-86.

Zhang, S., Wei, Z., Wang, Y. and Liao, T., 2018. Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary. *Future Generation Computer Systems*, *81*, pp.395-403.

Zhang, W., Li, Y. and Wang, S., 2019. Learning document representation via topic-enhanced LSTM model. *Knowledge-Based Systems*, *174,* pp. 194-204.

Zhang, X., Zhao, J. and LeCun, Y., 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems* (pp. 649-657).

Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H. and Xu, B., 2016. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. *arXiv preprint arXiv:1611.06639*.

Zikmund, W.G. and Babin, B.J., 2012. *Essentials of marketing research*. Nelson Education.

APPENDICES

A MyCustomerLens Company Profile

MyCustomerLens is a digital analytics company, which specialises in extracting insight from customer feedback. Its primary service is providing "*a real-time customer insight platform, which converts customer feedback into business intelligence*" so that their corporate business clients can make "*faster, more informed decisions*", as explained by their CEO, Paul Roberts.

"Our bespoke algorithms collect and analyse real-time comments from social media, surveys and feedback forms." -MyCustomerLens team

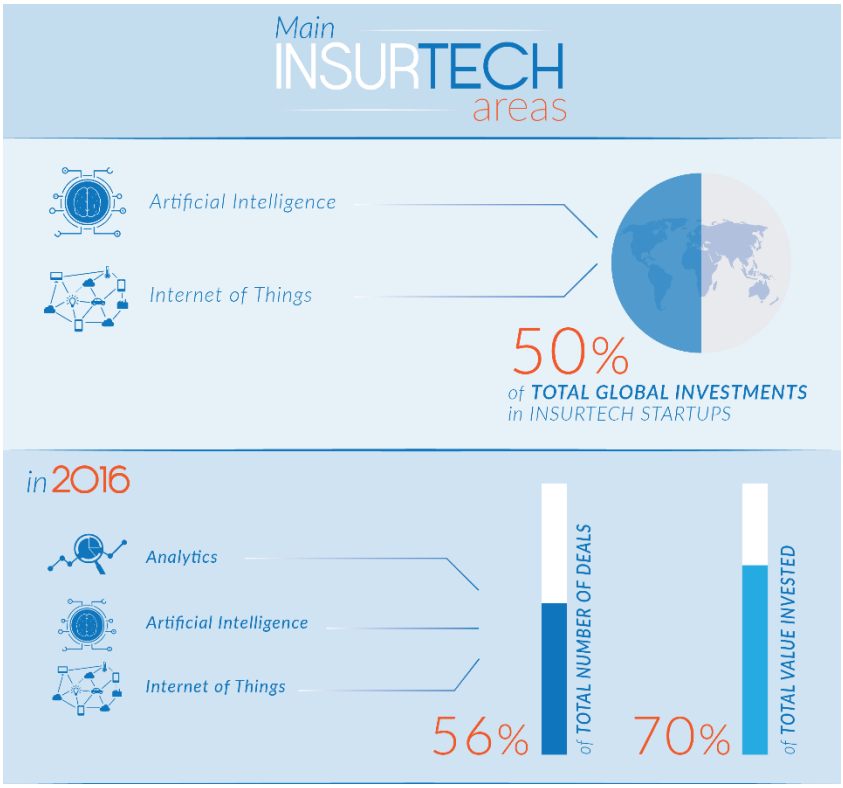
The company specialises in providing small businesses (currently primarily in the sports, health and leisure industry) a suite of solutions that are designed to improve the relationship between the companies and their consumers. MyCustomerLens currently commissions tools for automatic feedback collection from the web and social media websites, tools for analysing the collected textual data, as well as a web-based platform for account management, where their business clients can access data dashboards for with insights for their organisations.

Read more about MyCustomerLens here.

B The Insurance Industry in 2019: Market Overview

The insurance industry in 2019 has demonstrated sustained economic growth, continuously rising interest rates, and higher income from investments in the field, which is speculated to sustain growth in future years as well.

One of industry's key drivers is determined to be the growth of the technological field and software development that supports automation in the insurance sector, with some examples being blockchain technology, Internet of Things (IoT), cognitive applications and cloud computing. On-demand insurance has also emerged in recent years as a response to changes in consumer behaviour, with several applications have already occurred, sparking the field of InsurTech, which engages in real-time as-needed insurance coverage.



These changes provide immense opportunities for companies in the field; however, require strengthening of the relationship between companies and consumers, with this being understood by the leaders in the field, who have increased the percentage of investment in real-time data analytics tools. A key competitive advantage for insurance companies in 2019 is their agility, responsiveness to market changes and relationship marketing.

These insights have been summarised from Deloitte's (2019) insurance industry report, which can be read in full here.

C Systematic Literature Review Methodology Process

Table C-5.7-1 Systematic Review Worksheet, based on the PRISMA methodology (adapted from Moher et al., 2009)

	Database searched	Date of Search	Search Terms	Filters applied		# of Records retrieved	# of Records excluded after screening	# of Records included
Record identification and Screening	**DATABASES**			Year	Type of Publication (Domain-specific)			
	Elsevier (Science Direct)	03.03.19	topic modelling sentiment analysis short form text	2014 - 2019	• Review Articles • Research Articles	1,850	1,772	78
	Emerald Insight	04.06.19	topic modelling sentiment analysis short form text		Articles & Chapters	260	256	4
	IEEE Xplore	04.06.19	topic modelling sentiment analysis short form text		• Journals & Magazines • Early Access Articles	417	398	19
	OTHER SOURCES							
	Handsearching	ongoing						
	Reference List Searching	ongoing						12
	NO DUPLICATE STUDIES FOUND AMONGST THE EXAMINED SAMPLE							
Eligibility	**DATABSE-EXTRACTED STUDIES AFTER FULL-TEXT ASSESSED FOR ELIGIBILITY:**	**ELIGIBILITY CRITERIA APPLIED TO DATABSE-EXTRACTED STUDIES** • Journal rating 3* and above on the 2018 ABS Academic Journal Guide **OR** (if not present) Journal Impact Factor (GIF) above 3 points, based on Web of Science's ranking of 2017 • Written in English						77
Included	**TOTAL # OF STUDIES INCLUDED:**							

D Data Extraction Procedure: Limitations, Stats and Queries

Table D-5.7-2 Facebook Data Extraction Procedure, Key Stats and Limitations

Facebook API Logs, Associated Queries and Records per query

DATE	STATUS	QUERY	RECORDS
2019-07-13	Facebook graph API call	TruShieldIns	100
2019-07-13	Facebook graph API call	admiralUK	384
2019-07-13	Facebook graph API call	1stCentral	486
2019-07-13	Facebook graph API call	aviva	854
2019-07-13	Facebook graph API call	TheAAUK	324
2019-07-13	Facebook graph API call	TheAAUK	324
2019-07-13	Facebook graph API call	axa	1288
2019-07-13	Facebook graph API call	axa	1288
2019-07-13	Facebook graph API call	axa	1288
2019-07-13	Facebook graph API call	TheAAUK	324
2019-07-13	Facebook graph API call	TheAAUK	324
2019-07-13	Facebook graph API call	allianzinsuranceuk	190
2019-07-13	Facebook graph API call	CoveaInsurancePlc	194
2019-07-13	Facebook graph API call	AspenInsurance	135
2019-07-13	Facebook graph API call	AgeasGroup	191
2019-07-13	Facebook graph API call	358211851051470	183
2019-07-13	Facebook graph API call	LV	323
2019-07-13	Facebook graph API call	InsuranceDotCom	106
2019-07-13	Facebook graph API call	TotallySportsInsurance	489
2019-07-13	Facebook graph API call	insure4sureUK	120
2019-07-13	Facebook graph API call	BQIGroup	103
2019-07-13	Facebook graph API call	Ins.Onl	103

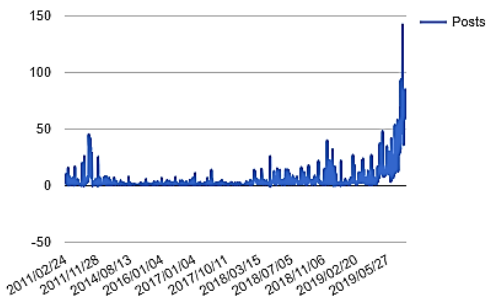
Extracted Data procedure Limitations

- Hourly extraction of up to 2,500 entries per query
- API: Facebook Graph
- Returns posts and replied from public Facebook groups, pages, events or profiles
- Returns up to 100 top level posts to/from a page, as well as up to 25 replies per post
- Replies to replies are not included

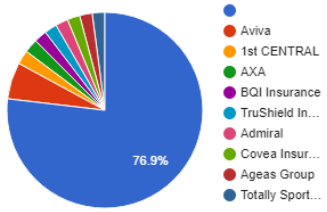
Overall Dataset Stats

Dataset Stats	
Dataset Name:	Insurance
Dataset Last Updated:	2019-07-13 17:18:11
Dataset Source:	facebook
Total Messages:	5573
Unique Posters:	38

Posts Distribution per date of Publication

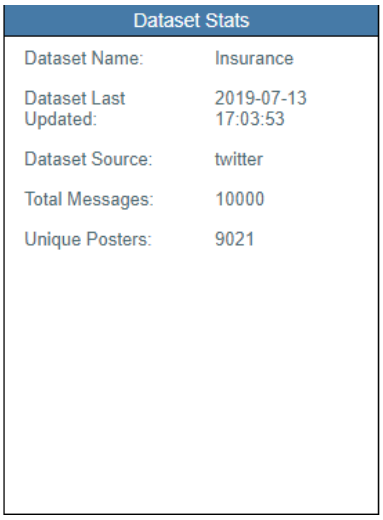
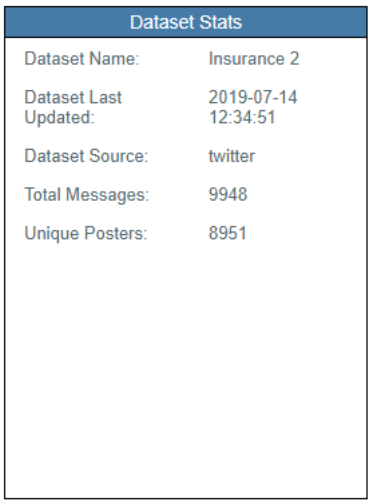
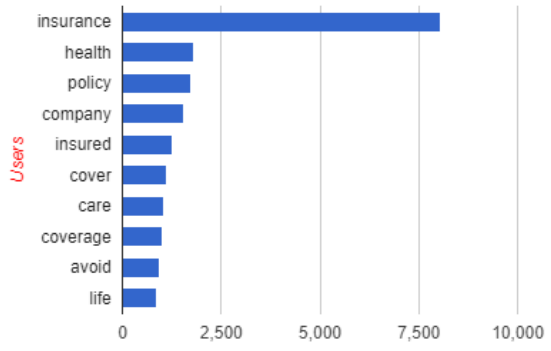
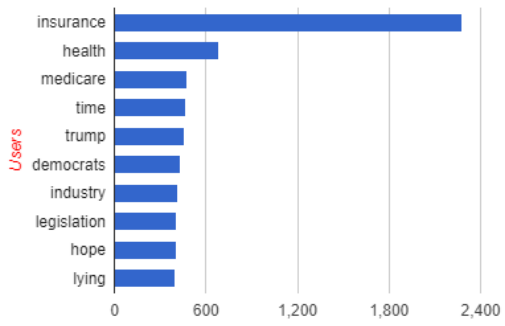


Posts Distribution per Publisher



Note: Publications from corporate entities from the Facebook dataset were manually identified and removed from the dataset.

Table D-5.7-3 Twitter Data Key Stats and Limitations

Stats	*Extracted Data procedure Limitations*
 	• API is Twitter REST API v1.1 search/tweets endpoint • Returns a collection of relevant tweets, matching the given query • To be noted that the search service is not an exhaustive source of tweets, as come might not be indexed • Typically, does not return tweets older than 1 week • Can retrieve up to 1000 most recent tweets
Most Frequent words	
	

E Ethical Approval Confirmation

ID	TITLE	STATUS	ACTION	LAST SUBMITTED	LAST REVIEWED
1000	Topic modelling, Sentiment analysis and Text classification of user-generated social media text, for identification of a user's stage in their Customer Journey	APPROVED	View	2019-08-07 00:30:48	2019-08-08 09:10:41

F Research Protocol

University of Strathclyde Glasgow

Computer and Information Sciences - local.cis

departmental information for staff and students

Home | Events | PGR | Safety | Systems Support | Teaching | Utilities

Browse:Home / Utilities / CIS Ethics Approval System

CIS Ethics Approval System

You are Lazarina Stoyanova (IMIP2018)

Return to Main

Application ID: 1000

Title of research:
Topic modelling, Sentiment analysis and Text classification of user-generated social media text, for identification of a user's stage in their Customer Journey

Summary of research (short overview of the background and aims of this study):
The purpose of this research is to create a system prototype that can read user-generated text from social media, extract its topic, sentiment polarity and classify the text automatically as one of 5 categories, which represent a stage in the customer journey process (i.e. expectation/awareness, consideration, purchase, retention, advocacy). The context of the user-generated text is insurance.

This survey's aim is to provide a basis for evaluation of the performance of machine learning models, which have been developed as part of the research. Specifically, you will be asked to read user-generated social media text data from Twitter or Facebook and perform the same tasks as the models, which have been developed as part of the study. The results will later be compared with the model's output as part of the analysis of the project.

How will participants be recruited?
Through Social Media Community Groups, Direct Communication and Snowballing

What will the participants be told about the proposed research study? Either upload or include a copy of the briefing notes issued to participants. In particular this should include details of yourself, the context of the study and an overview of the data that you plan to collect, your supervisor, and contact details for the Departmental Ethics Committee.
PDF File: None.
[copy of briefing notes]

Participation Information& Consent Form Welcome to the testing page for the prototype system I have been working on as part of my Master thesis research project. The title of the study is Topic modelling, Sentiment analysis and Text classification of user-generated social media text, for identification of a user's stage in their Customer Journey, and the research is being carried out in the Computer and Information Sciences Department of the University of Strathclyde.

The following few paragraphs will explain the purpose, method and aims of the study, as well as give some context for you - the participant, how data will be used and what is the aim of your contribution in the system development process. At the end, you will be asked to provide consent to how your information will be processed, after which you can start the survey!

What is the purpose of the research?
The purpose of this research is to create a system prototype that can read user-generated text from social media, extract its topic, sentiment polarity and classify the text automatically as one of 5 categories, which represent a stage in the customer journey process (i.e. expectation/awareness, consideration, purchase, retention, advocacy). The context of the user-generated text is insurance.

What is the aim of this survey and what will you be asked to do?
This survey's aim is to provide a basis for evaluation of the performance of machine learning models, which have been developed as part of the research. Specifically, you will be asked to read user-generated social media text data from Twitter or Facebook and perform the same tasks as the models, which have been developed as part of the study. The results will later be compared with the model's output as part of the analysis of the project.

Specifically, after reading the user-generated text, you will be asked to assess its sentiment polarity, identify the topics that the user has discussed, and identify the user's customer journey stage, with options being: expectation/awareness, consideration, purchase, retention, advocacy. You will also be able to indicate if you think the text is unrelated to a user's insurance purchase customer journey.

What is required of you when completing the survey?

Nothing else but to read each question instruction carefully and answer the questions honestly, providing your opinion regarding the performance of presented models.

How long will this survey take to complete?
The study should take you around 7-10 minutes to complete.

When can you complete this survey?
The survey will only be available for 3 days from the 9.08-11.08.2019 (inclusive), following which the survey will be taken down.

Who can complete this study?
Anyone that has command over the English language can take part in the system evaluation process.

Are there any risks involved in taking part?
Since this research processes user-generated text, some of the examples provided will display the expression of negative language and negative advocacy. Although all effort has been placed in reducing the impact this has on participants, such examples are included for performance evaluation purposes.

If you are uncomfortable with negative advocacy, it is recommended that you do not take part in the study.

Do you have to take part?
Your participation in this research is voluntary. You have the right to withdraw at any point during the study, for any reason, and without any prejudice. Once the answers to the survey have been completed, withdraws are no longer possible.

What data is being collected from this survey?
The survey will collect opinion data, but will not at any stage ask for personally-identifiable information, e.g. you name, date of birth, address, etc.

You will only be asked to identify your age, as this will help highlight future opportunities in expanding the study population by including members of various demographic categories. A visualisation of this data will also be included in the written thesis.

The data collection and handling protocol complies with GDPR (2018) and UK's ESRC (2015) ethical guidelines for researchers.

How will your response data be analysed and why?
Collected responses will be analysed using statistical tests to ensure a scientific research approach. The aim of the human-agent evaluation of a system is to identify areas of poor performance of the machine learning models, which can be highlighted for system improvement in subsequent stages of development.

Where will the information be stored and how long will it be kept for?
During the study, research data will be kept securely on Qualtrics' survey platform and will be accessed by the researcher alone following a secure authentication process. Upon completion of the research project, the researcher will store the data on a cloud platform, with it being kept in a password-protected process, ensuring two-factor authentication.

The University of Strathclyde can request data for validation purposes.

Who should you contact in the event that you want to discuss the study further?
If you would like to contact either myself (the researcher) or the Chief Investigator of this study to discuss any elements of this research, please send emails to:
-lazarina.stoyanova.2014@uni.strath.ac.uk for Lazarina Stoyanova - Researcher
-william.wallace@strath.ac.uk for William Wallace - Research Supervisor (Chief Investigator)

This research was granted ethical approval by the University of Strathclyde Ethics Committee. If you have any questions or concerns, during or after the research, or wish to contact an independent person in association with this research, please contact:

Secretary of Departmental Ethics Committee
Department of Computer and Information Sciences,
Livingstone Tower
Richmond Street
Glasgow
G1 1XH

or send an email to ethics@cis.strath.ac.uk

What happens next?
Nothing is required from you upon completion of the survey. If upon completion of the survey, you would like to share it with anyone, you are more than welcome to do so, using the link that has been provided to you.

If you are happy to participate, click the button below and get started with the model evaluation procedures. If upon reading the information sheet, you have decided to withdraw your participation, you can exit the study now. In both cases I humbly thank you for your time.

By clicking the button below, you acknowledge that:
- your participation in the study is voluntary;
- you are aware that you may choose to terminate your participation in the study at any time and for any reason, prior to submitting your responses;
- you agree that you submitted responses will be analysed and used as described above
- a summary of all responses will be attached as an appendix for the written project thesis, as well as in any written publications outwith the University of Strathclyde, if

such are to arise following assessment of the academic rigour of this study

Please note that this survey will be best displayed on a laptop or desktop computer. Some features may be less compatible for use on a mobile device.

How will consent be demonstrated? Either upload or include here a copy of the consent form/instructions issued to participants. It is particularly important that you make the rights of the participants to freely withdraw from the study at any point (if they begin to feel stressed for example), nor feel under any pressure or obligation to complete the study, answer any particular question, or undertake any particular task. Their rights regarding associated data collected should also be made explicit.
PDF File: None.
The [above] form will be displayed at the start of the survey, with participants being asked to choose either of these two options [listed below]

I consent, begin the study
I do not consent, I do not wish to participate

What will participants be expected to do? Either upload or include a copy of the instructions issued to participants along with a copy of or link to the survey, interview script or task description you intend to carry out. Please also confirm (where appropriate) that your supervisor has seen and approved both your planned study and this associated ethics application.
PDF File: None.
PDF File: None.
[link to survey]
https://strathbusiness.qualtrics.com/jfe/form/SV_a9U7XK7sezuSehL

Supervisor has approved the type of survey and questions asked.

What data will be collected and how will it be captured and stored? In particular indicate how adherence to the Data Protection Act and the General Data Protection Regulation (GDPR) will be guaranteed and how participant confidentiality will be handled.
The survey will collect opinion data, but will not at any stage ask for personally-identifiable information, e.g. you name, date of birth, address, etc.

You will only be asked to identify your age, as this will help highlight future opportunities in expanding the study population by including members of various demographic categories. A visualisation of this data will also be included in the written thesis.

The data collection and handling protocol complies with GDPR (2018) and UK's ESRC (2015) ethical guidelines for researchers.

How will the data be processed? (e.g. analysed, reported, visualised, integrated with other data, etc.) Please pay particular attention to describing how personal or sensitive data will be handled and how GDPR regulations will be met.
No personal data will be collected or stored in association with this survey.

Collected responses will be analysed using statistical tests to ensure a scientific research approach. The aim of the human-agent evaluation of a system is to identify areas of poor performance of the machine learning models, which can be highlighted for system improvement in subsequent stages of development.
The data from this survey will be integrated with other data, generated as part of the research, and later triangulated with data from previously published studies.

How and when will data be disposed of? Either upload a copy of your data management plan or describe how data will be disposed.
PDF File: None.
During the study, research data will be kept securely on Qualtrics' survey platform and will be accessed by the researcher alone following a secure authentication process. Upon completion of the research project, the researcher will store the data on a cloud platform, with it being kept in a password-protected process, ensuring two-factor authentication.

G Electronic Consent Form

Default Report

Topic Modelling, Sentiment Analysis and Text Classification to Identify Stages of the Customer Journey

August 10, 2019 10:45 PM BST

Q1 - Participation Information& Consent Form

Welcome to the testing page for the prototype system I have been working on as part of my Master thesis research project.

The title of the study is Topic modelling, Sentiment analysis and Text classification of user-generated social media text, for identification of a user's stage in their Customer Journey, and the research is being carried out in the Computer and Information Sciences Department of the University of Strathclyde.

The following few paragraphs will explain the purpose, method and aims of the study, as well as give some context for you - the participant, how data will be used and what is the aim of your contribution in the system development process. At the end, you will be asked to provide consent to how your information will be processed, after which you can start the survey!

What is the purpose of the research?

The purpose of this research is to create a system prototype that can read user-generated text from social media, extract its topic, sentiment polarity and classify the text automatically as one of 5 categories, which represent a stage in the customer journey process (i.e. expectation/awareness, consideration, purchase, retention, advocacy). The context of the user-generated text is insurance.

What is the aim of this survey and what will you be asked to do?

This survey's aim is to provide a basis for evaluation of the

performance of machine learning models, which have been developed as part of the research. Specifically, you will be asked to read user-generated social media text data from Twitter or Facebook and perform the same tasks as the models, which have been developed as part of the study. The results will later be compared with the model's output as part of the analysis of the project. Specifically, after reading the user-generated text, you will be asked to assess its sentiment polarity, identify the topics that the user has discussed, and identify the user's customer journey stage, with options being: expectation/awareness, consideration, purchase, retention, advocacy. You will also be able to indicate if you think the text is unrelated to a user's insurance purchase customer journey.

What is required of you when completing the survey?

Nothing else but to read each question instruction carefully and answer the questions honestly, providing your opinion regarding the performance of presented models.

How long will this survey take to complete?

The study should take you around 7-10 minutes to complete.

When can you complete this survey?

The survey will only be available for 3 days from the 9.08-11.08.2019 (inclusive), following which the survey will be taken down.

Who can complete this study?

Anyone that has command over the English language can take part in the system evaluation process.

Are there any risks involved in taking part?

Since this research processes user-generated text, some of the examples provided will display the expression

of negative language and negative advocacy. Although all effort has been placed in reducing the impact this has on participants, such examples are included for performance evaluation purposes. If you are uncomfortable with negative advocacy, it is recommended that you do not take part in the study.

Do you have to take part?

Your participation in this research is voluntary. You have the right to withdraw at any point during the study, for any reason, and without any prejudice. Once the answers to the survey have been completed, withdraws are no longer possible.

What data is being collected from this survey?

The survey will collect opinion data, but will not at any stage ask for personally-identifiable information, e.g. you name, date of birth, address, etc. You will only be asked to identify your age, as this will help highlight future opportunities in expanding the study population by including members of various demographic categories. A visualisation of this data will also be included in the written thesis. The data collection and handling protocol complies with GDPR (2018) and UK's ESRC (2015) ethical guidelines for researchers.

How will your response data be analysed and why?

Collected responses will be analysed using statistical tests to ensure a scientific research approach. The aim of the human-agent evaluation of a system is to identify areas of poor performance of the machine learning models, which can be highlighted for system improvement in subsequent stages of development.

Where will the information be stored and how long will it be kept for?

During

the study, research data will be kept securely on Qualtrics' survey platform and will be accessed by the researcher alone following a secure authentication process. Upon completion of the research project, the researcher will store the data on a cloud platform, with it being kept in a password-protected process, ensuring two-factor authentication. The University of Strathclyde can request data for validation purposes.

Who should you contact in the event that you want to discuss the study further?

If you would like to contact either myself (the researcher) or the Chief Investigator of this study to discuss any elements of this research, please send emails to:

- lazarina.stoyanova.2014@uni.strath.ac.uk for Lazarina Stoyanova – Researcher

-william.wallace@strath.ac.uk for William Wallace - Research Supervisor (Chief Investigator)

This research was granted ethical approval by the University of Strathclyde Ethics Committee.

If you have any questions or concerns, during or after the research, or wish to contact an independent person in association with this research, please contact:

Secretary of Departmental Ethics Committee

Department of Computer and Information Sciences,

Livingstone Tower

Richmond Street

Glasgow

G1 1XH or send an email to ethics@cis.strath.ac.uk. ----------

---------- What happens next?
Nothing is required from you upon completion of the

survey. If upon completion of the survey, you would like to share it with anyone, you are

more than welcome to do so, using the link that has been provided to you. If you are happy to participate, click the button below and get started with the model evaluation procedures.

If upon reading the information sheet, you have decided to withdraw your participation, you can exit the study now. In both cases I humbly thank you for your time.

By clicking the button below, you acknowledge that:

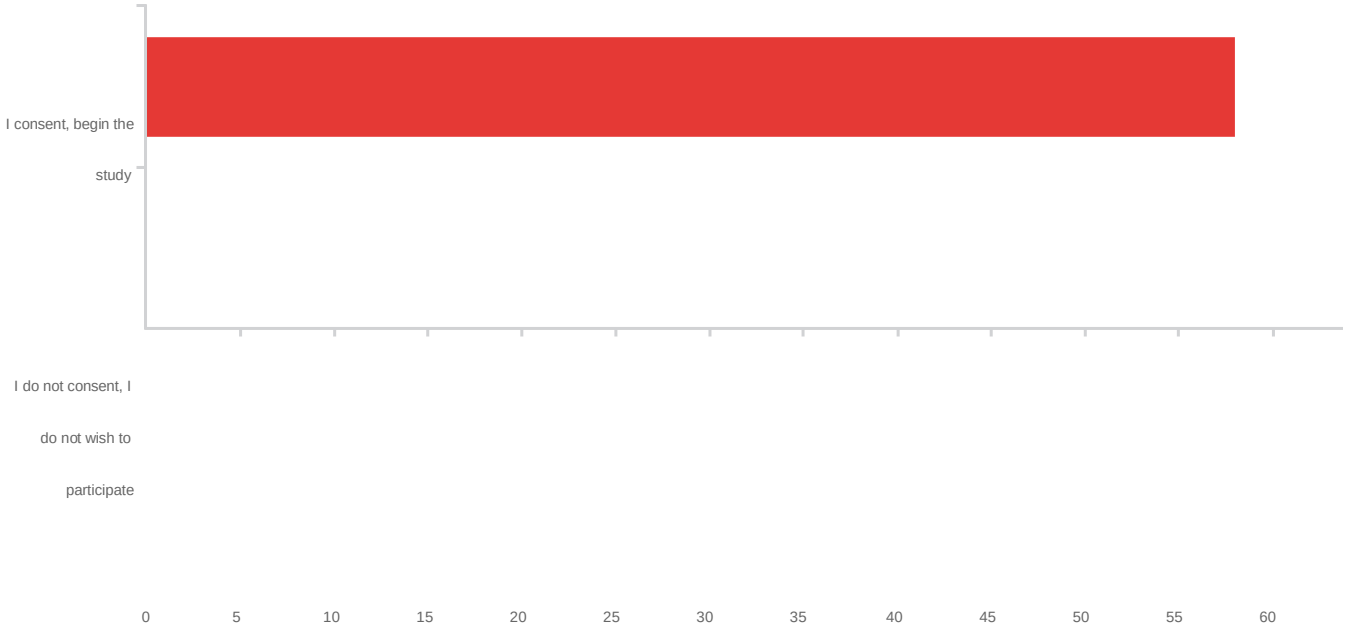
- your participation in the study is voluntary;

- you are aware that you may choose to terminate your participation in the study at any time and for any reason, prior to submitting your responses;

- you agree that you submitted responses will be analysed and used as described above

- a summary of all responses will be attached as an appendix for the written project thesis, as well as in any written publications outwith the University of Strathclyde, if such are to arise following assessment of the academic rigour of this study

Please note that this survey will be best displayed on a laptop or desktop computer. Some features may be less compatible for use on a mobile device.


I consent, begin the study
I do not consent, I do not wish to participate
0
5
10
15
20
25
30
35
40
45
50
55
60

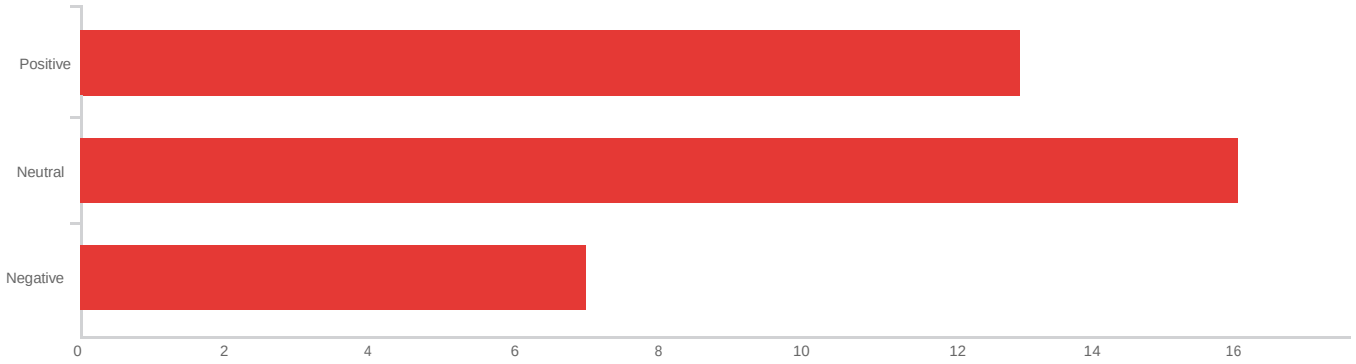
#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	Participation Information& Consent Form	1.00	1.00	1.00	0.00	0.00	58

#	Field		Choice Count
1	I consent, begin the study	100.00%	58
2	I do not consent, I do not wish to participate	0.00%	0
			58

Showing rows 1 - 3 of 3

H Survey Response Data

Q5 - For the following text: In the worst case of flooding, I hope to get a She Shed with my insurance money. #HurricaneBarry2019 #HurricaneBarry How would you rate the sentiment expressed in this text?



#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	For the following text: In the worst case of flooding, I hope to get a She Shed with my insurance money. #HurricaneBarry2019 #HurricaneBarry How would you rate the sentiment expressed in this text?	1.00	3.00	1.83	0.73	0.53	36

#	Field		Choice Count
1	Positive	36.11%	13
2	Neutral	44.44%	16
3	Negative	19.44%	7
			36

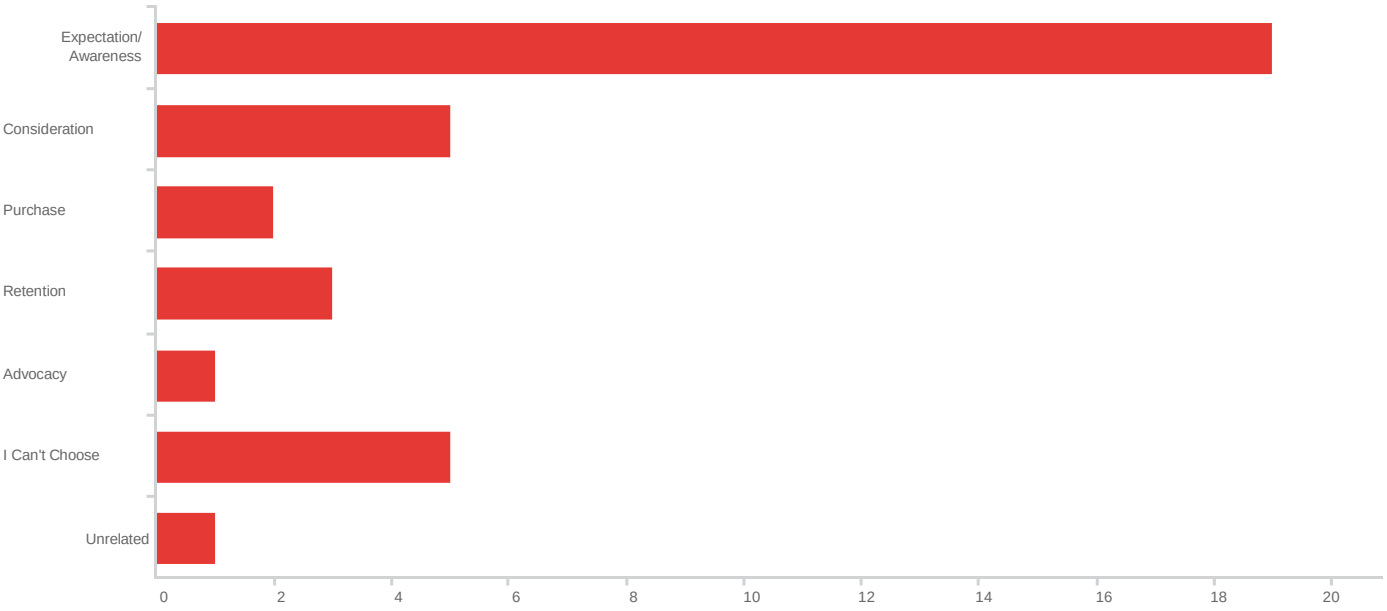
Showing rows 1 - 4 of 4

Q7 - What topics do you consider being the key topics of this tweet? Please indicate between 3-5 words, separated by commas (,).

What topics do you consider being the key topics of this tweet? Please indi...
flood, hurricane barry, need of shelter
Flooding, insurance, she shed
money, humour, attention
flooding, insurance, #HurricaneBarry
Hurricane, flooding, insurance
Hurricanebarry, she shed, insurance money
Hope to get money
hurricane, flooding, She Shed, insurance money
Hurricane Barry, flooding, insurance
government policy about critical situations
Flood, hurricane,insurance,money
Floading, hurricane bary, insurance money
Hurricane, insurance, flooding
Beware, fearful, practical
Insurance, flooding, hurricane
disaster, fear, insurance, money
3
Flooding, Insurance, money
flood , insurance , hurricane
Hurricane, insurance, hope
Insurance, flooding, hurricane

What topics do you consider being the key topics of this tweet? Please indi...
insurance,scam,disaster
Insurance,flooding,hurricane
worst case, insurance, flooding, hurricane
insurance, money, begging
Hurricane Barry, Insurance, Flooding,
Safety
hurricanes, flooding, insurance, claim
Claiming,Insurance,money
Hurricane Barry 2019
Flooding insurance my money
Hurricane, insurance, money

Q8 - At which stage of the customer journey of purchasing insurance cover would you classify the user, who has posted this text?

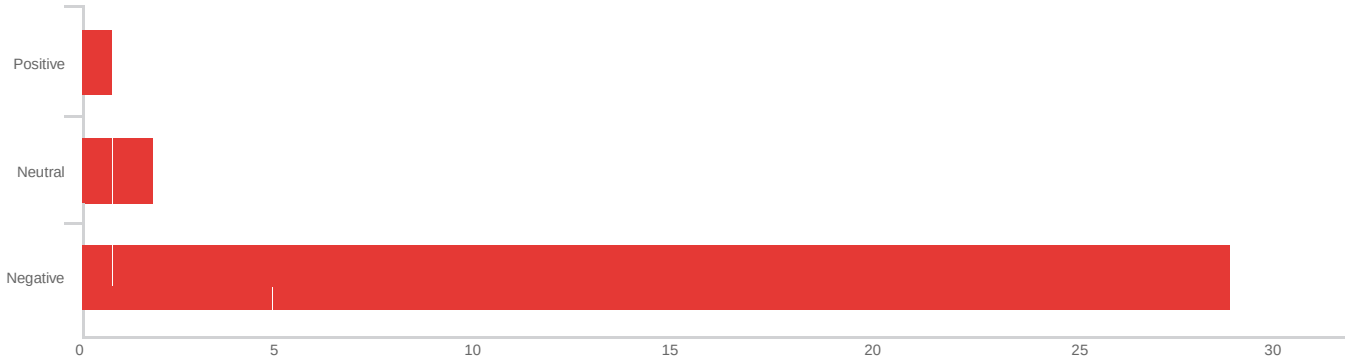


#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	At which stage of the customer journey of purchasing insurance cover would you classify the user, who has posted this text?	1.00	7.00	2.47	1.96	3.86	36

#	Field		Choice Count
1	Expectation/ Awareness	52.78%	19
2	Consideration	13.89%	5
3	Purchase	5.56%	2
4	Retention	8.33%	3
5	Advocacy	2.78%	1
6	I Can't Choose	13.89%	5
7	Unrelated	2.78%	1
			36

Showing rows 1 - 8 of 8

Q17 - For the following text: @COMPANY Worst insurance company I have ever seen. As per my experience they don't provide the offered insurance amount . It's a trap for the customers. They just loot the people. I don't get how @USER has allowed such companies to operate their business. How would you rate the sentiment expressed in this text?



#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	For the following text: @COMPANY Worst insurance company I have ever seen. As per my experience they don't provide the offered insurance amount . It's a trap for the customers. They just loot the people. I don't get how @USER has allowed such companies to operate their business. How would you rate the sentiment expressed in this text?	1.00	3.00	2.88	0.41	0.17	32

#	Field		Choice Count
1	Positive	3.13%	1
2	Neutral	6.25%	2
3	Negative	90.63%	29
			32

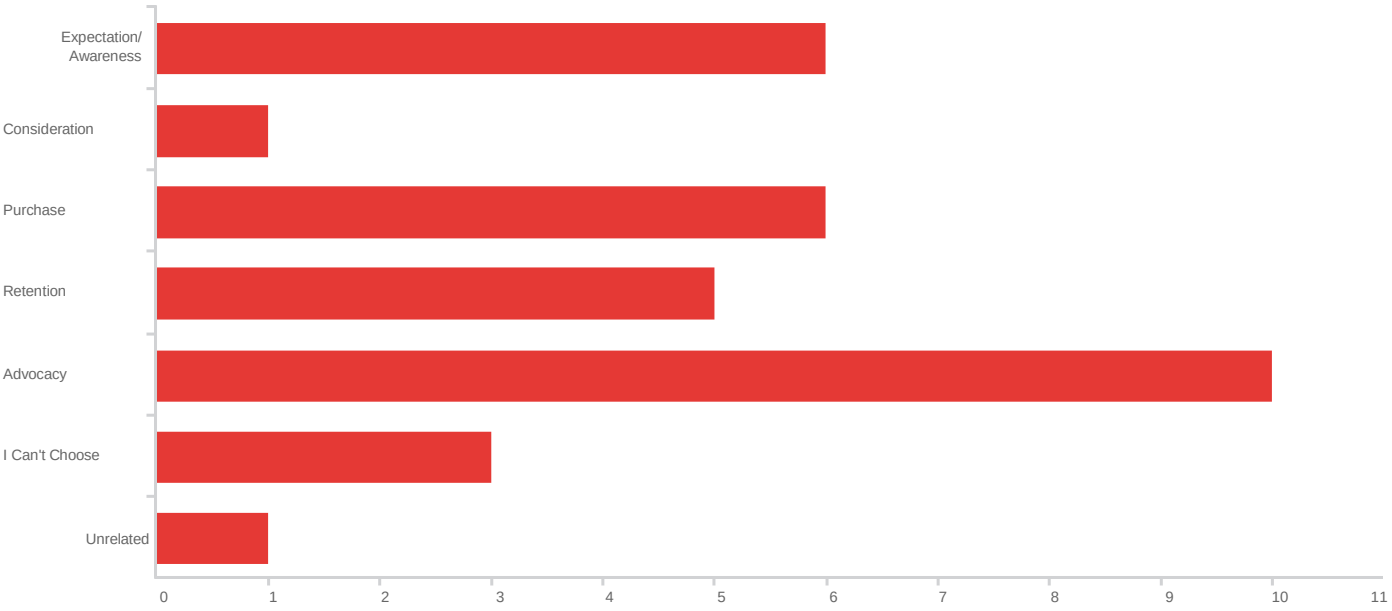
Showing rows 1 - 4 of 4

Q18 - What topics do you consider being the key topics of this tweet? Please indicate between 3-5 words, separated by commas (,).

What topics do you consider being the key topics of this tweet? Please indi...
Insurance, business, worst
poor service, raising awareness, complaint
insurance company, worst, loot
Loot, worst, insurance
Insurance, experience, worst, trap
insurance, company, worst
insurance company, insurance money, bad experience with insurance company
problematic insurance service
Insurance, disappointment, fraud
Bad service, complaint
Worst insurance, don’t provide, trap, loot
Insurance, fraud, @company
competition, bad advertising, not enough information
Worst, companies, experience
Worst, insurance, amount, loot
insurance , misleading , company
Denial, bad company, unsatisfied customer
@company, insurance, feedback
insurance,scam,fraud
Insurance,business,experience,bad
worst, my experience, trap

What topics do you consider being the key topics of this tweet? Please indi...
insurance company, trap, disappointment
Insurance, Complaint, Customer Feedback,
Insurance scam, unfair politics
complaint, insurance, regulation
Anger,insurance,company,lied
Jipped by insurance
Worst insurance company
Trap for customers, insurance amount, worst company

Q19 - At which stage of the customer journey of purchasing insurance cover would you classify the user, who has posted this text?

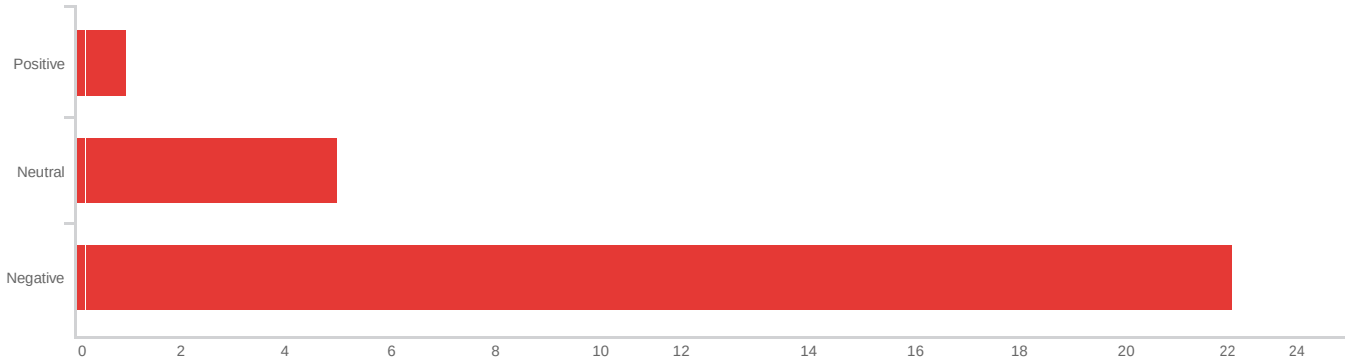


#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	At which stage of the customer journey of purchasing insurance cover would you classify the user, who has posted this text?	1.00	7.00	3.78	1.71	2.92	32

#	Field	Choice Count
1	Expectation/ Awareness	18.75% 6
2	Consideration	3.13% 1
3	Purchase	18.75% 6
4	Retention	15.63% 5
5	Advocacy	31.25% 10
6	I Can't Choose	9.38% 3
7	Unrelated	3.13% 1
		32

Showing rows 1 - 8 of 8

Q26 - For the following text: @COMPANY @USER We are trying our best to get a life. To get a roof-to get walls-to get the insurance to return our calls. All the while being told to 'get over it'. How would you rate the sentiment expressed in this text?



#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	For the following text: @COMPANY @USER We are trying our best to get a life. To get a roof-to get walls-to get the insurance to return our calls. All the while being told to 'get over it'. How would you rate the sentiment expressed in this text?	1.00	3.00	2.75	0.51	0.26	28

#	Field		Choice Count
1	Positive	3.57%	1
2	Neutral	17.86%	5
3	Negative	78.57%	22
			28

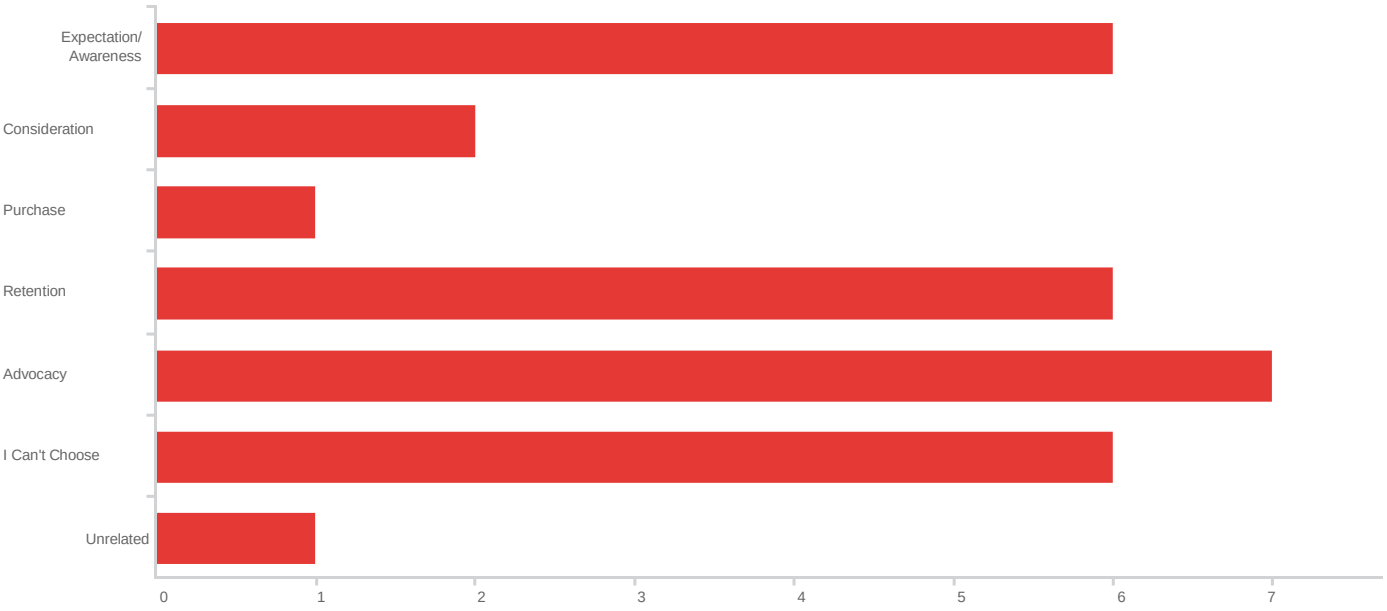
Showing rows 1 - 4 of 4

Q27 - What topics do you consider being the key topics of this tweet? Please indicate between 3-5 words, separated by commas (,).

What topics do you consider being the key topics of this tweet? Please indi...
Roof, insurance, answer
coping, complaints
Get, Over, It
Insurance, life, roof, walls
insurance, insurance company
insurance company abusing the customer rights
Insurance, desparate, help
Struggling to get coverage
Get a life, get over it
Insurance, communication, @company
disappointment
Life, insurance, trying.
Life, return, get, over, it
insurance , unavailable , bad
poverty,disappointment,service
Insurance,bad,experience,company
negativity, company, user, insurance
disappointment, expectations, insurance
Customer complaint, problem, Insurance, frustration
Dissapointment
claim, insurance, complaint

What topics do you consider being the key topics of this tweet? Please indi...
Criticised,while,they,try,their,best
Unfair insurance practices
Life, insurance, return calls

Q28 - At which stage of the customer journey of purchasing insurance cover would you classify the user, who has posted this text?

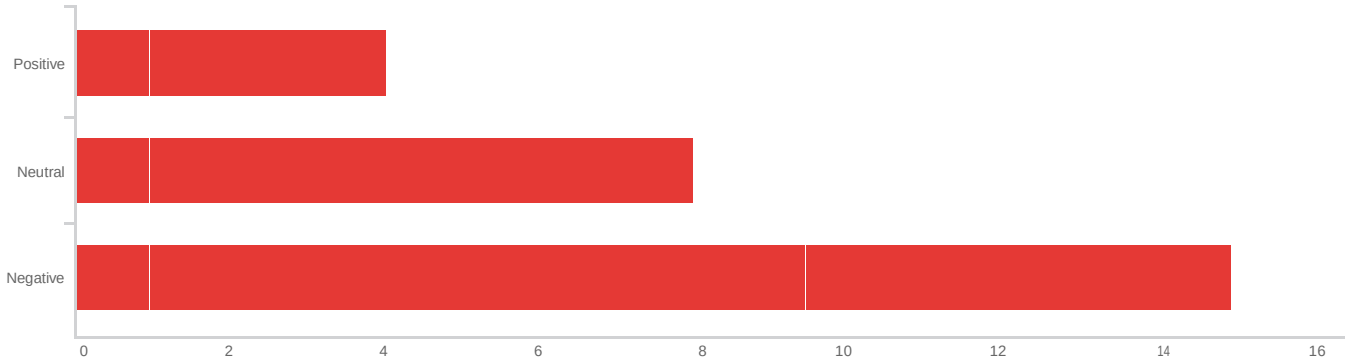


#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	At which stage of the customer journey of purchasing insurance cover would you classify the user, who has posted this text?	1.00	7.00	3.97	1.88	3.55	29

#	Field	Choice Count
1	Expectation/ Awareness	20.69% 6
2	Consideration	6.90% 2
3	Purchase	3.45% 1
4	Retention	20.69% 6
5	Advocacy	24.14% 7
6	I Can't Choose	20.69% 6
7	Unrelated	3.45% 1
		29

Showing rows 1 - 8 of 8

Q31 - For the following text: @USER Thankfully her kids aren't in school yet and the insurance company finally gave her a rental. But yeah, the complication of it all is a pain in the ass. And she's JUST back. How would you rate the sentiment expressed in this text?



#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	For the following text: @USER Thankfully her kids aren't in school yet and the insurance company finally gave her a rental. But yeah, the complication of it all is a pain in the ass. And she's JUST back. How would you rate the sentiment expressed in this text?	1.00	3.00	2.41	0.73	0.54	27

#	Field	Choice Count
1	Positive	14.81% 4
2	Neutral	29.63% 8
3	Negative	55.56% 15
		27

Showing rows 1 - 4 of 4

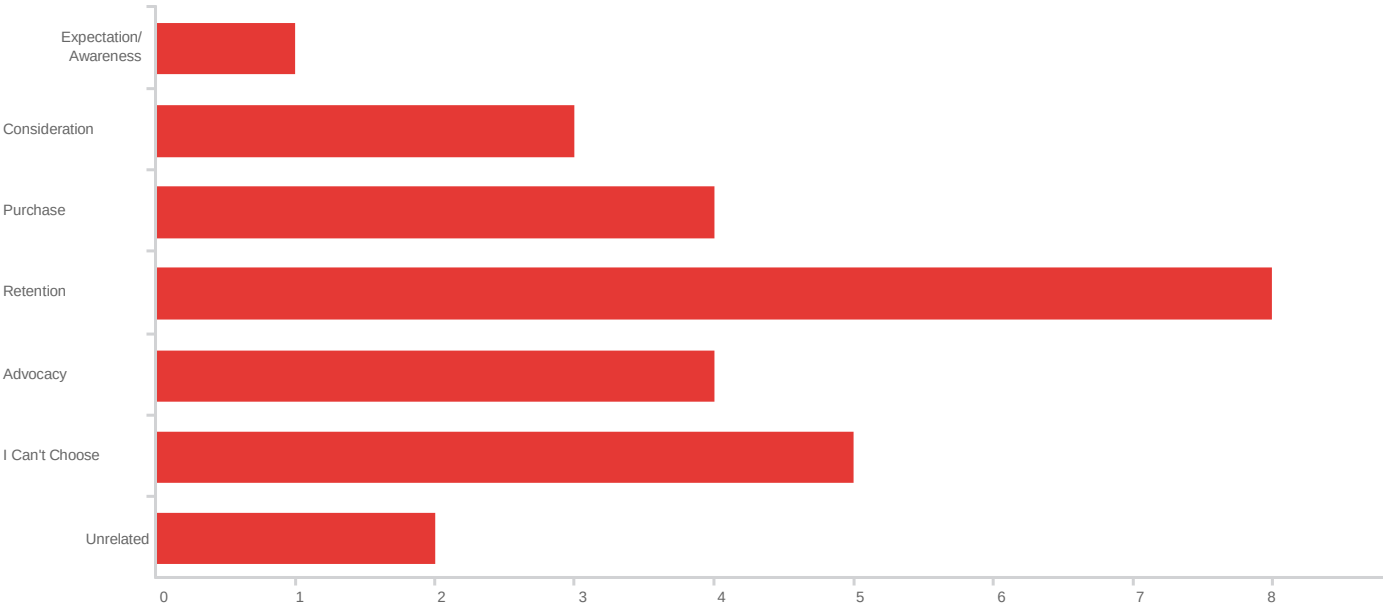
Q30 - What topics do you consider being the key topics of this tweet? Please indicate between 3-5 words, separated by commas (,).

What topics do you consider being the key topics of this tweet? Please indi...
Kids; insurance company; complication
discontent, opinion, real situations
Finally, complication, just
Insurance, kids, rental
insurance company
Car rental, work, school kids
Pain, in, the, ass
Insurance, rental, complication
constatation
Complication,kids, pain.
Ass, complication, pain
insurance , complication , company
insurance cover,paid,thanks
Insurance,experience, relief
complication, children, rental, insurance
rental, complications, could have been better
Insurance claim,
complaint, insurance, claim
Frustrated,over,complications
Insurance helps out a family
Thankfully

What topics do you consider being the key topics of this tweet? Please indi...

Rental, company, insurance

Q32 - At which stage of the customer journey of purchasing insurance cover would you classify the user, who has posted this text?

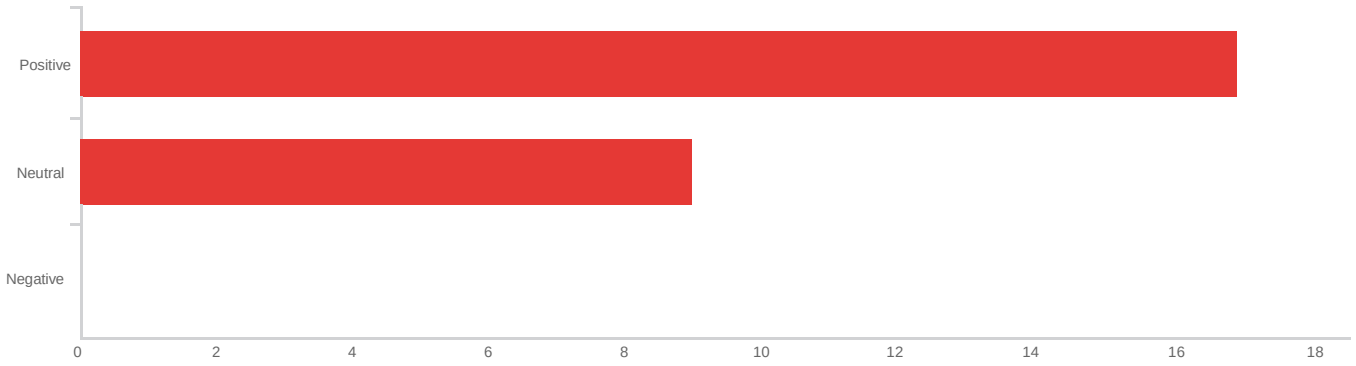


#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	At which stage of the customer journey of purchasing insurance cover would you classify the user, who has posted this text?	1.00	7.00	4.26	1.55	2.41	27

#	Field	Choice Count
1	Expectation/ Awareness	3.70% 1
2	Consideration	11.11% 3
3	Purchase	14.81% 4
4	Retention	29.63% 8
5	Advocacy	14.81% 4
6	I Can't Choose	18.52% 5
7	Unrelated	7.41% 2
		27

Showing rows 1 - 8 of 8

Q22 - For the following text: Check out the easiest and quickest way to find affordable coverage with us. It only takes a few minutes to compare the best quotes from a variety of providers, giving you the most choice when it comes to finding your home insurance policy. [link] How would you rate the sentiment expressed in this text?



#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	For the following text: Check out the easiest and quickest way to find affordable coverage with us. It only takes a few minutes to compare the best quotes from a variety of providers, giving you the most choice when it comes to finding your home insurance policy. [link] How would you rate the sentiment expressed in this text?	1.00	2.00	1.35	0.48	0.23	26

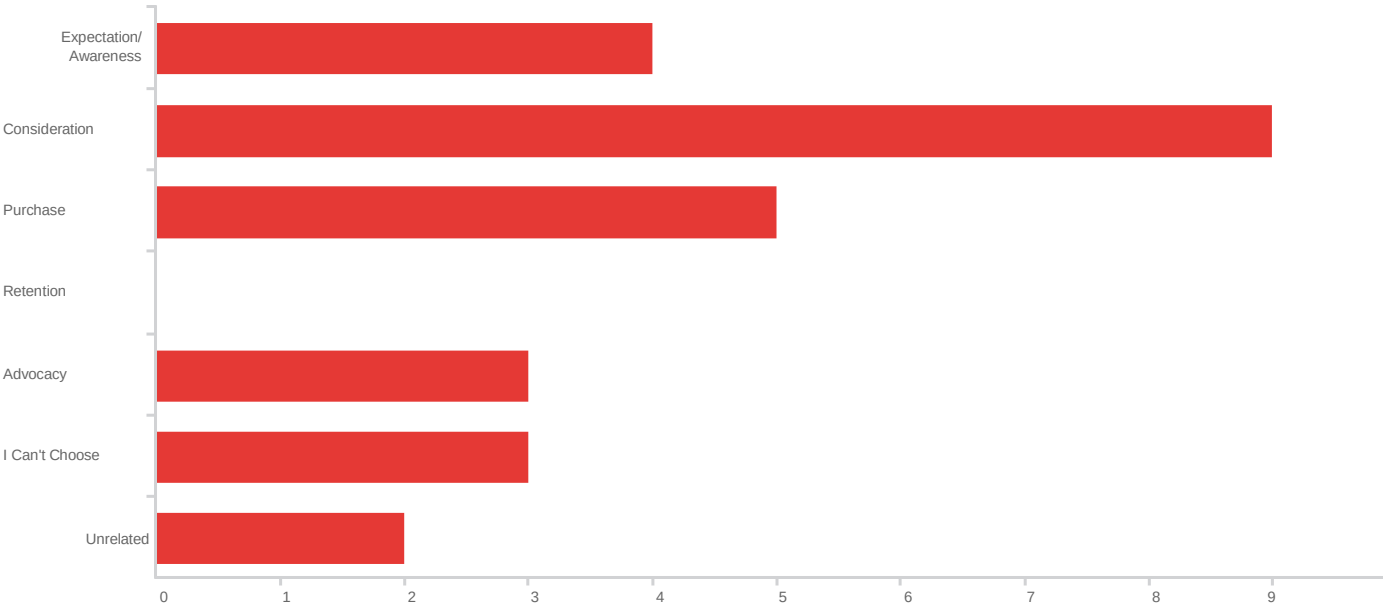
#	Field		Choice Count
1	Positive	65.38%	17
2	Neutral	34.62%	9
3	Negative	0.00%	0
			26

Showing rows 1 - 4 of 4

Q23 - What topics do you consider being the key topics of this tweet? Please indicate between 3-5 words, separated by commas (,).

What topics do you consider being the key topics of this tweet? Please indi...
advertisement, advice, information
Easiest, variety, compare
Quickest, affordable, best, choice
insurance commercial
Find coverage, best deal
Affordable, coverage, variety, most, choice
Home insurance policy, promotion, affordability
Home, compare, insurance.
choice , insurance , advertisement
advert,easy,quick
Insurance,sell, choice,home
affordable plans, home, insurance, policy
offer, advertisement, choices
Home Insuracne, selling, positive,
Promoting a product
Helping,people,find,best,insurance,easily
Short and sweet insurance pitch
Most choice, easiest quickest
Quotes, providers, choice, home insurance

Q24 - At which stage of the customer journey of purchasing insurance cover would you classify the user, who has posted this text?

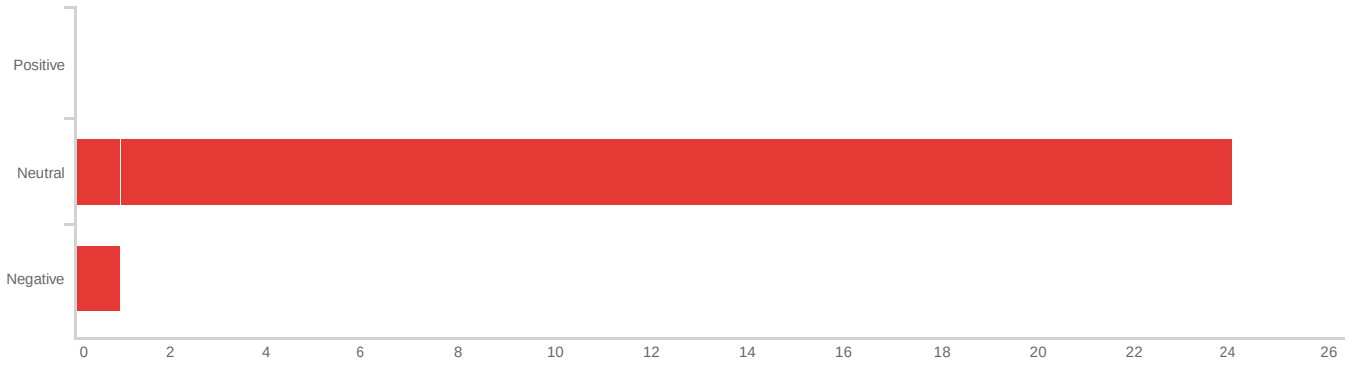


#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	At which stage of the customer journey of purchasing insurance cover would you classify the user, who has posted this text?	1.00	7.00	3.23	1.91	3.64	26

#	Field	Choice Count
1	Expectation/ Awareness	15.38% 4
2	Consideration	34.62% 9
3	Purchase	19.23% 5
4	Retention	0.00% 0
5	Advocacy	11.54% 3
6	I Can't Choose	11.54% 3
7	Unrelated	7.69% 2
		26

Showing rows 1 - 8 of 8

Q34 - For the following text: Does @COMPANY have an accident policy? I know @COMPANY does and you are insured as a rider. As a rider with @COMPANY, are you insured in the case of an accident? How would you rate the sentiment expressed in this text?



#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	For the following text: Does @COMPANY have an accident policy? I know @COMPANY does and you are insured as a rider. As a rider with @COMPANY, are you insured in the case of an accident? How would you rate the sentiment expressed in this text?	2.00	3.00	2.04	0.20	0.04	25

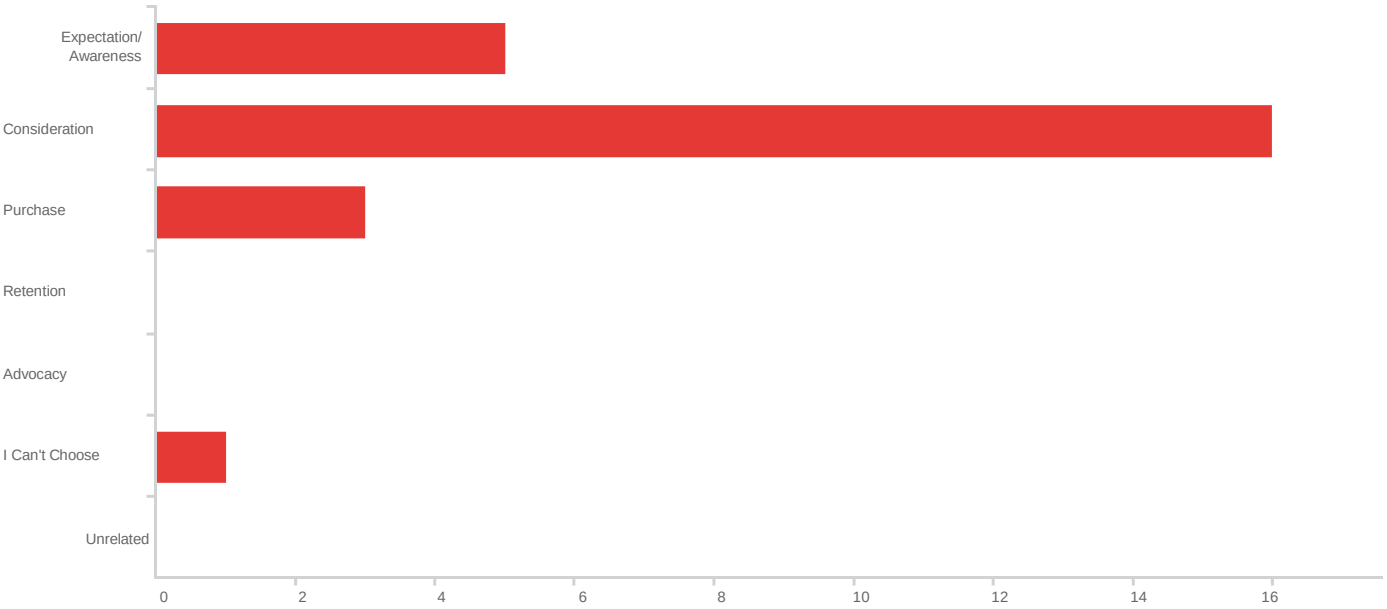
#	Field	Choice Count
1	Positive	0.00% 0
2	Neutral	96.00% 24
3	Negative	4.00% 1
		25

Showing rows 1 - 4 of 4

Q35 - What topics do you consider being the key topics of this tweet? Please indicate between 3-5 words, separated by commas (,).

What topics do you consider being the key topics of this tweet? Please indi...
accidents, drivers, questions
Accident, Are you, policy
Accident, rider, insured
Query, accident coverage
Accident, policy, rider
Insurance, accident policy, rider insurance, @company
Q&A
Rider, accident, insured.
question , insurance , policy
question,insurance,cover
Insurance,policy,accident,query
company insurance, accident, company rider
question, lack of information, comparing
Question, Vehicle Insurance, Customer question,
insurance, claim, query
Questions,about,accident,insurance,policies
Comparing insurance coverage
Rider, policy, accident

Q36 - At which stage of the customer journey of purchasing insurance cover would you classify the user, who has posted this text?

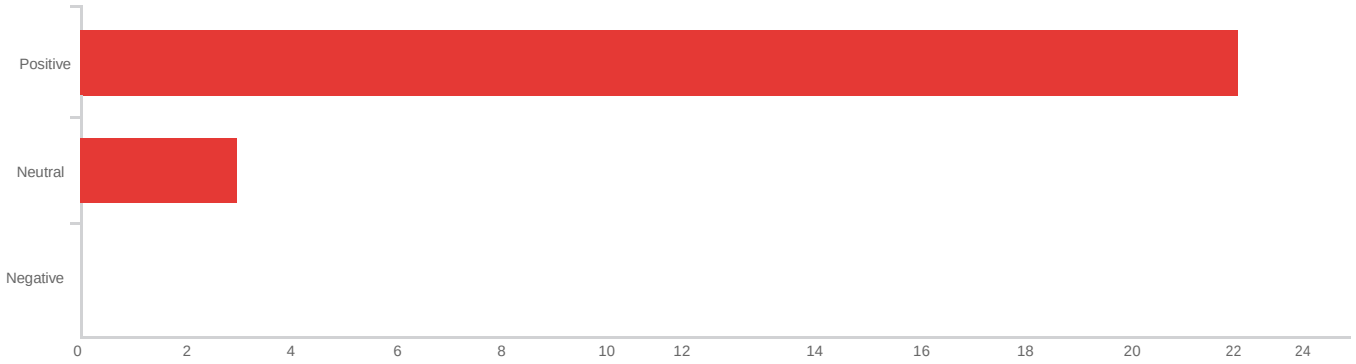


#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	At which stage of the customer journey of purchasing insurance cover would you classify the user, who has posted this text?	1.00	6.00	2.08	0.98	0.95	25

#	Field	Choice Count
1	Expectation/ Awareness	20.00% 5
2	Consideration	64.00% 16
3	Purchase	12.00% 3
4	Retention	0.00% 0
5	Advocacy	0.00% 0
6	I Can't Choose	4.00% 1
7	Unrelated	0.00% 0
		25

Showing rows 1 - 8 of 8

Q40 - For the following text: The wife got her dream car today (aside from a G-Wagon) Insured by @COMPANY #BIGCoverage How would you rate the sentiment expressed in this text?



#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	For the following text: The wife got her dream car today (aside from a G-Wagon) Insured by @COMPANY #BIGCoverage How would you rate the sentiment expressed in this text?	1.00	2.00	1.12	0.32	0.11	25

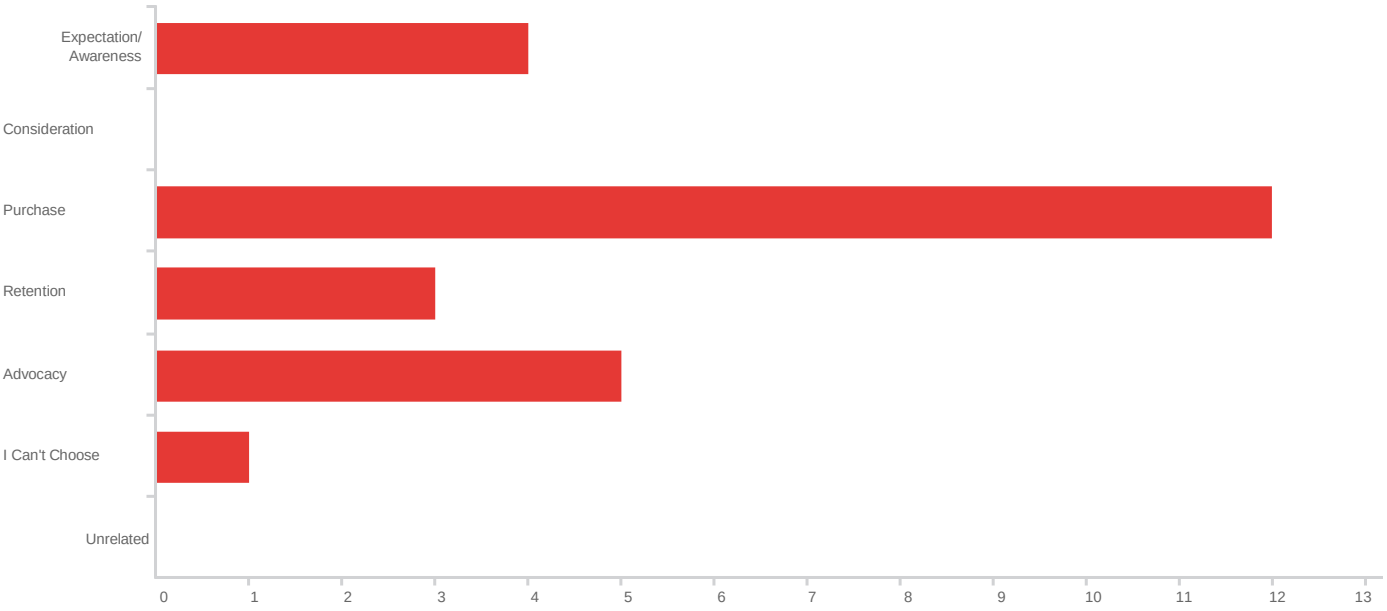
#	Field		Choice Count
1	Positive	88.00%	22
2	Neutral	12.00%	3
3	Negative	0.00%	0
			25

Showing rows 1 - 4 of 4

Q41 - What topics do you consider being the key topics of this tweet? Please indicate between 3-5 words, separated by commas (,).

What topics do you consider being the key topics of this tweet? Please indi...
purchase, gifts, companionship
Insured, dream, #
Wife, dream, insured
New car, good insurance policy
Dream, big, car
Insurance, car purchase, coverage, @company
Pleasure
Dream, car, insured.
car , insurance , company
dream,insurance,coverage
Insured,insurance, purchase,car
dream car, insured, wife
coverage, buy, car
Car Insurance, Happy Customer, Customer Feedback,
Satisfaction
boast, insurance
Happy,with,car,and,insurance
Insurance covers luxes
Dream car
Coverage, insurance, today

Q42 - At which stage of the customer journey of purchasing insurance cover would you classify the user, who has posted this text?

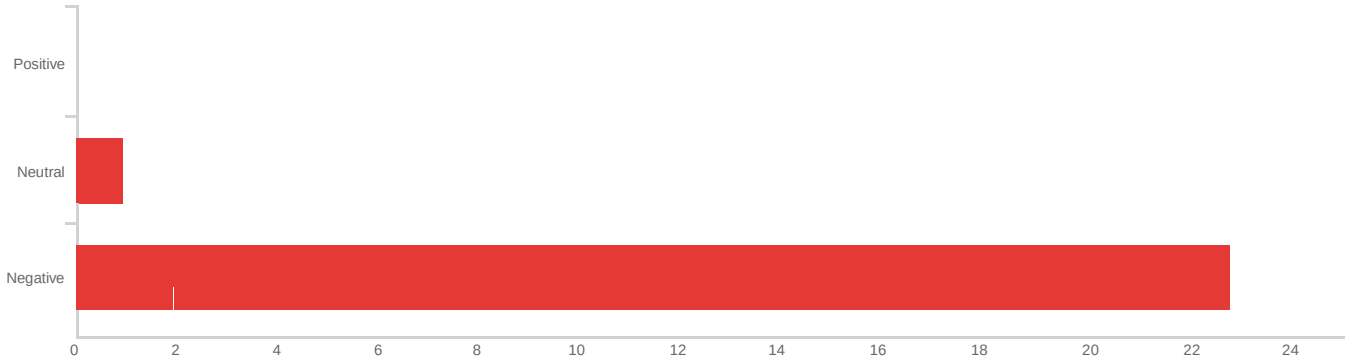


#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	At which stage of the customer journey of purchasing insurance cover would you classify the user, who has posted this text?	1.00	6.00	3.32	1.35	1.82	25

#	Field		Choice Count
1	Expectation/ Awareness	16.00%	4
2	Consideration	0.00%	0
3	Purchase	48.00%	12
4	Retention	12.00%	3
5	Advocacy	20.00%	5
6	I Can't Choose	4.00%	1
7	Unrelated	0.00%	0
			25

Showing rows 1 - 8 of 8

Q43 - For the following text: @COMPANY avoid at all costs. This place is a joke. Your insured member was at fault and caused an accident involving 3 other cars two weeks ago. Countless calls and nobody has returned my call regarding my damaged vehicle. Your claims adjuster will not return calls. How would you rate the sentiment expressed in this text?



#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	For the following text: @COMPANY avoid at all costs. This place is a joke. Your insured member was at fault and caused an accident involving 3 other cars two weeks ago. Countless calls and nobody has returned my call regarding my damaged vehicle. Your claims adjuster will not return calls. How would you rate the sentiment expressed in this text?	2.00	3.00	2.96	0.20	0.04	24

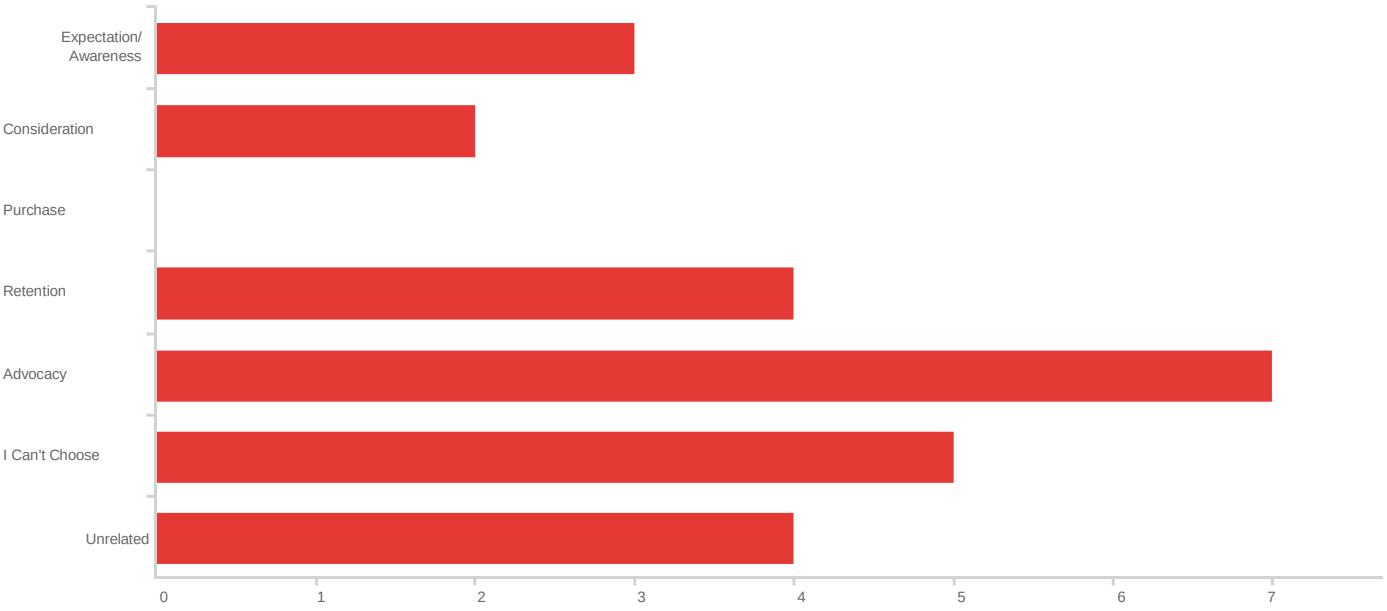
#	Field		Choice Count
1	Positive	0.00%	0
2	Neutral	4.17%	1
3	Negative	95.83%	23
			24

Showing rows 1 - 4 of 4

Q44 - What topics do you consider being the key topics of this tweet? Please indicate between 3-5 words, separated by commas (,).

What topics do you consider being the key topics of this tweet? Please indi...
discontent, conflict, complaint
Avoid, joke, fault
Avoid, joke, fault, accident
Poor response, accident
Avoid, costs, countless
Insurance, @company, customer-claims adjuster communication
Not , calls, damaged, accident.
vehicle , accident , insurance , claims
poor service,scam,claim
Insurance,experience, user
fedup, useless service, no return calls
disappointment, calls, joke
Complaint, Car Insurance, Frustration,
insurance, customer service, claim, complaint
Angry,over,company,ignorance
Terrible customer service
Avoid will not return calls
Claims, return calls, accident

Q45 - At which stage of the customer journey of purchasing insurance cover would you classify the user, who has posted this text?

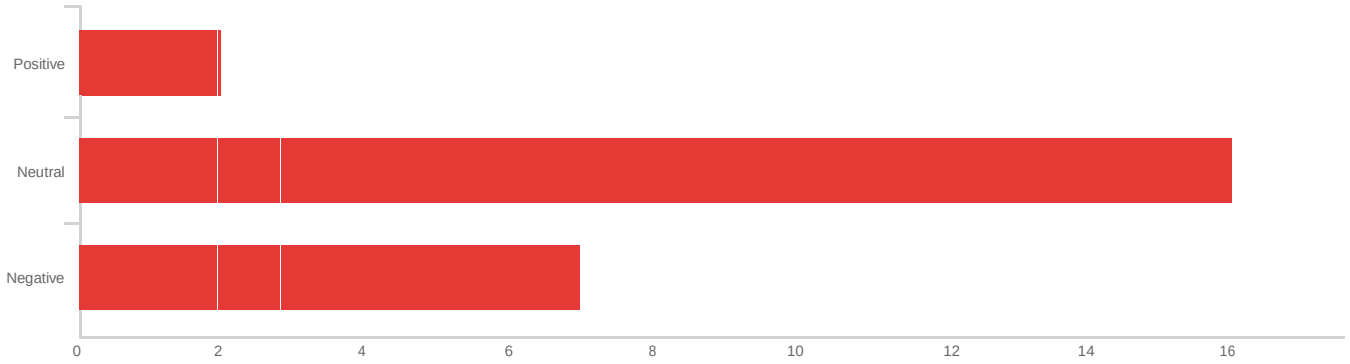


#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	At which stage of the customer journey of purchasing insurance cover would you classify the user, who has posted this text?	1.00	7.00	4.64	1.87	3.51	25

#	Field	Choice Count
1	Expectation/ Awareness	12.00% 3
2	Consideration	8.00% 2
3	Purchase	0.00% 0
4	Retention	16.00% 4
5	Advocacy	28.00% 7
6	I Can't Choose	20.00% 5
7	Unrelated	16.00% 4
		25

Showing rows 1 - 8 of 8

Q37 - For the following text: @USER A feedlot I work with just told me yesterday insurance won't allow them to put plastic on silage pile this fall because of worker safety How would you rate the sentiment expressed in this text?



#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	For the following text: @USER A feedlot I work with just told me yesterday insurance won't allow them to put plastic on silage pile this fall because of worker safety How would you rate the sentiment expressed in this text?	1.00	3.00	2.20	0.57	0.32	25

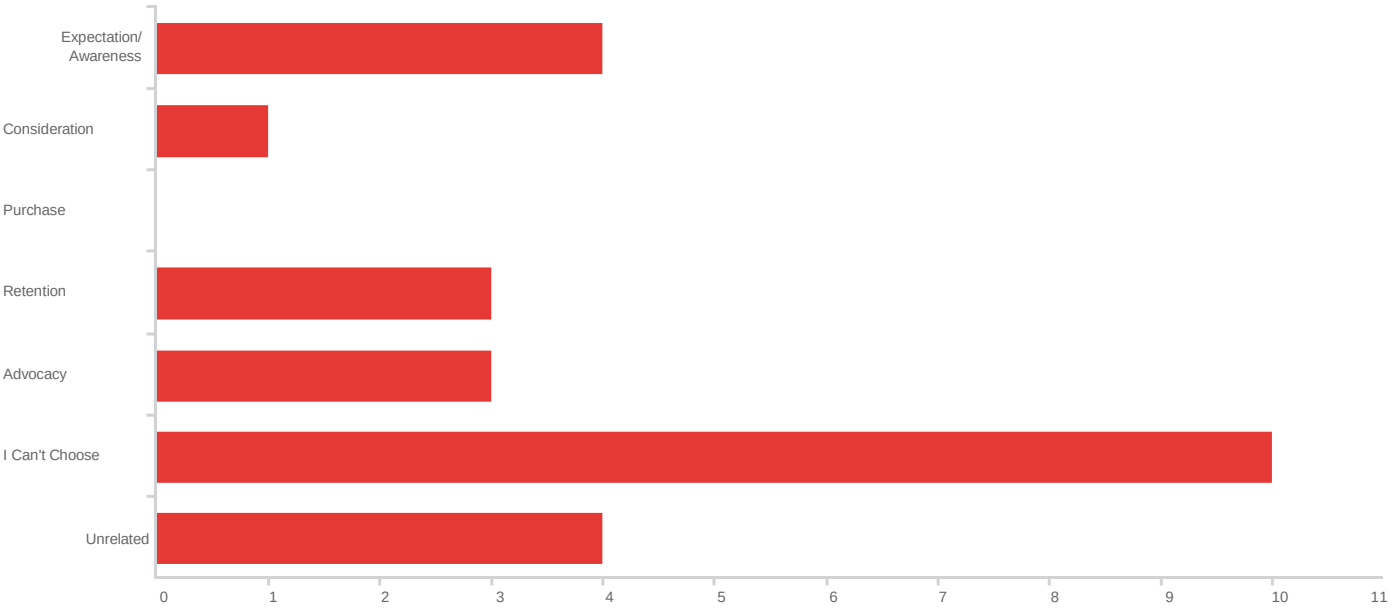
#	Field		Choice Count
1	Positive	8.00%	2
2	Neutral	64.00%	16
3	Negative	28.00%	7
			25

Showing rows 1 - 4 of 4

Q38 - What topics do you consider being the key topics of this tweet? Please indicate between 3-5 words, separated by commas (,).

What topics do you consider being the key topics of this tweet? Please indi...
policy, rules, update
Won't allow, safety
Insurance, feedlot, safety
Barrier at work, plastic
Won’t, allow, plastic
Insurance, worker safety policy
disappointment
Insurance, worker, safety,
insurance , safety , workers
change,safety,worker
Insurance,user,policy
safety first, insurance
information, reason, safety
health and safety, insurance, workplace safety
Informative,on,worker,safety,insurance
Insurance extends preventative caution
Won't allow them
Sileage, worker safety, plastic

Q39 - At which stage of the customer journey of purchasing insurance cover would you classify the user, who has posted this text?



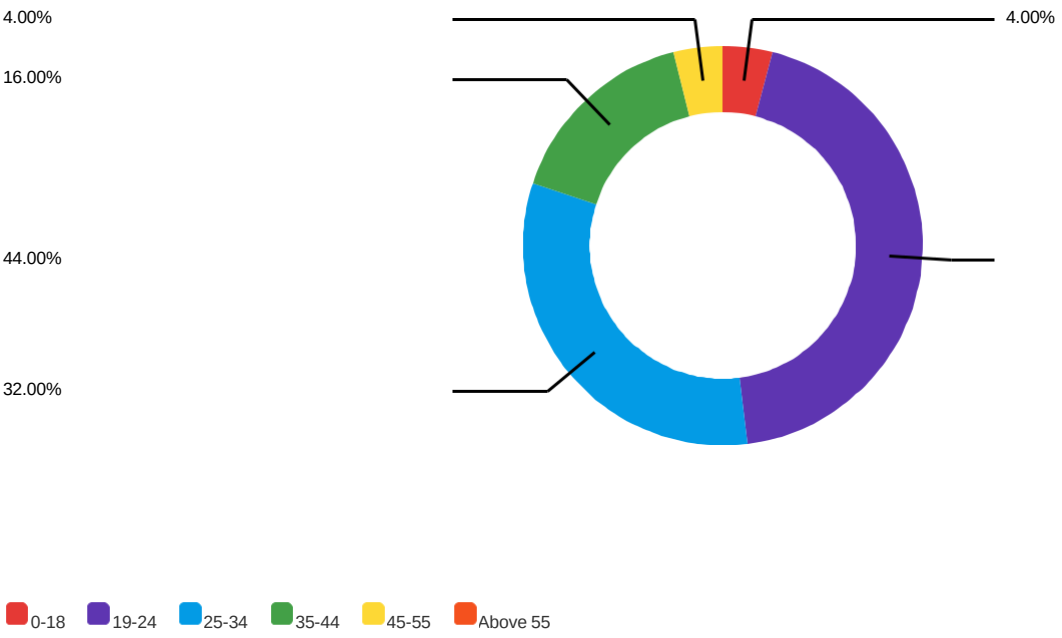
#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	At which stage of the customer journey of purchasing insurance cover would you classify the user, who has posted this text?	1.00	7.00	4.84	2.01	4.05	25

#	Field	Choice Count
1	Expectation/ Awareness	16.00% 4
2	Consideration	4.00% 1
3	Purchase	0.00% 0
4	Retention	12.00% 3
5	Advocacy	12.00% 3
6	I Can't Choose	40.00% 10
7	Unrelated	16.00% 4
		25

Showing rows 1 - 8 of 8

Q41 - Thank you for your responses so far! As indicated at the start of the survey, we would like to collect some demographic data for our survey participants. Please do indicate your age group, so we can analyse the data and use that improve our future research sample targeting.

#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	Thank you for your responses so far! As indicated at the start of the survey, we would like to collect some demographic data for our survey participants. Please do indicate your age group, so we can analyse the data and use that improve our future research sample targeting.	1.00	5.00	2.72	0.92	0.84	25

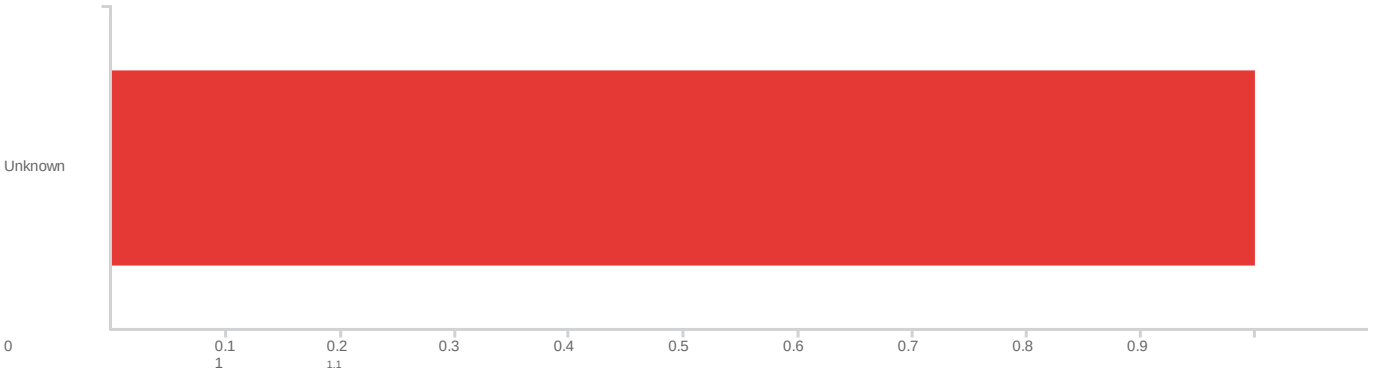


#	Field	Choice Count
1	0-18	4.00% 1
2	19-24	44.00% 11
3	25-34	32.00% 8

#	Field		Choice Count
4	35-44	16.00%	4
5	45-55	4.00%	1
6	Above 55	0.00%	0
			25

Showing rows 1 - 7 of 7

Q35 - Topics



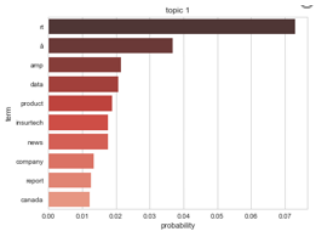
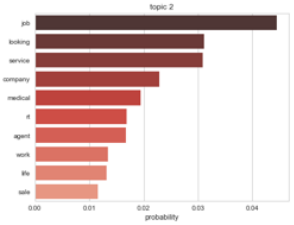
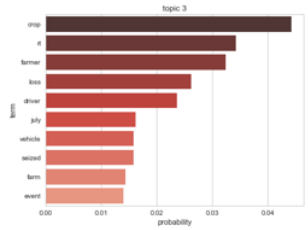
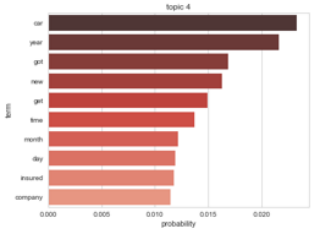
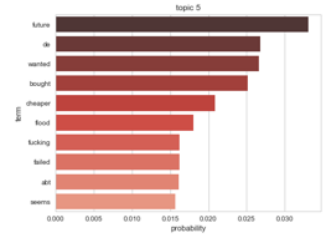
#	Field	Choice Count
1	Unknown	100.00% 1

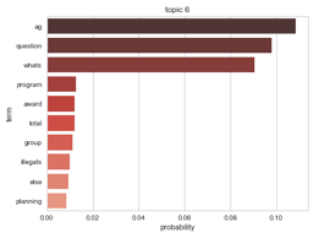
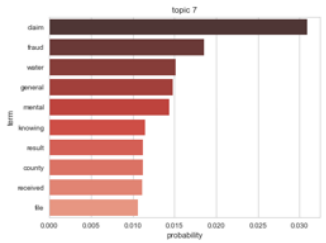
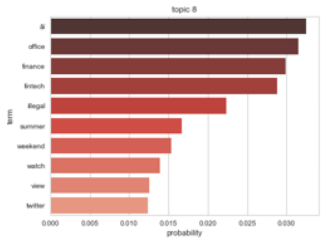
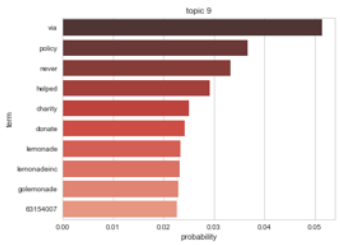
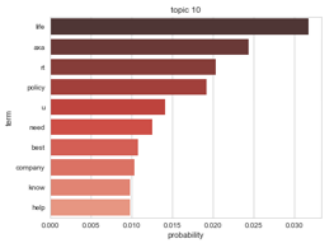
Showing rows 1 - 1 of 1

End of Report

I Location Map of Study Participants

J LDA Topic Model Term Probability Demonstration on Selected Texts

Topic #	LDA topic weight	Description	Topic Term Probability Bar chart
1	0.505	rt â amp data product	
2	0.149	job looking service company medical	
3	0.16	crop rt farmer loss driver	
4	0.39	car year got new get	
5	0.13	future de wanted bought cheaper	

6	0.322	ag question whats program award	
7	0.094	claim fraud water general mental	
8	0.144	âï office finance fintech illegal	
9	0.175	via policy never helped charity	
10	0.109	life axa rt policy u	

K Annex Documentation: Supporting Python Code, Extracted Data and Survey Data

Required installations prior to running Python files

- UMAP-learn,
- textblob,
- genism,
- xgboost,
- pyLDAvis

Download and install:

- Glove.6b word2vec – Glove.6B.100d
- Wikinews300d1mvec

Python Files (uncompiled) attached in **CODE** (folder):

- **Mypreprocessing&sentimentanalysis.py** – Python code for all data cleaning, feature extraction and data exploration procedures, including sentiment analysis classifiers
- **myLSA.py –** Python code for LSA model
- **myLDA.py –** Python code for LDA model
- **mytextclassification.py –** Python code for all supervised shallow and deep learner approaches for text classification
- **myKmeansfulldataset.py –** Python code for Kmeans clustering on full dataset
- **myKmeanstestdataset.py –** Python code for Kmeans clustering on test dataset (100 samples)
- **myNBtextclassification.py –** Naïve Bayes for text classification

In the **surveydata** (folder):

- **topicwordclouds.py** – file for generating word clouds from study participant topic models