



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

Text analytics of customer service social media interactions using Machine Learning and Artificial Intelligence

Anthony Mwirikia Njonge B.Sc. M.Sc.

*A report submitted to the University of Wales in partial fulfilment of the requirements
for the degree of Master of Science Software Engineering (Web Development)*

Faculty of Architecture, Computing and Engineering
Wales Institute of Science and Art
University of Wales Trinity Saint David

January 2020

Declaration

I, Anthony Mwirikia Njonge declare that I am the sole author of this project; that all references cited have been consulted; that I have conducted all work of which this is a record, and that the finished work lies within the prescribed word limits.

This work has not previously been accepted as part of any other degree submission.

Signed



Date 31st of January 2020

Form of Consent

I, Anthony Mwirikia Njonge hereby consent that my project, submitted in candidature for the M.Sc. Software Engineering degree, if successful, may be made available for inter-library loan or photocopying (subject to the law of copyright), and that the title and abstract may be made available to outside organisations.

Signed



Date 31st of January 2020

Acknowledgements

I would like to give thanks to *Ngai* for the all that I am. I know it is all about Him, and the blessing of his Son as a gift to the world, and His Spirit for direction.

I would like to thank my dear wife, Edna. You are my best gift from the Almighty and I am glad to have you in my life. I know *bila wewe ngingekuwa tu pale pale bila mwelekeo*. I love you.

To my lovely children, Jesse, Kyra and Liam. Thank you for being very patient with me as Dad did his thing. Thank you for allowing me time to disappear to the library or university those many days. I hope my journey will inspire you to greatness and to follow your dreams. I love you more than words can describe.

To my Mummy, you are the greatest mother any son would wish for.

To my siblings and their families, thanks for always being there for me. You are always in my heart.

I would like to express my sincere gratitude to my supervisor, Gordon Dickers, for his guidance, help and direction in this dissertation project. His immense patience and knowledge in the field helped me gain confidence and knowledge. To the UWTSD SOAC lecturers thank you for getting me back to the passion of computing.

To my fellow colleagues in the DVLA M.Sc. cohort. Thank you for making this such a great experience. I know it would have been less so without you.

To the DVLA team, Craig, Michelle, Helen, and the ETLAB thank you.

Abstract

Social media is becoming a growing avenue for customer service. As government moves to deliver more services in the cloud, it is therefore important that customer service provisions match this drive. It is therefore an onus on organisations to develop channels to engage their customers wherever they are. At present this is being done through the use of chatbots, social media and internet interactive channels. However, this requires new ways of getting insight of the issues being experience by the customers. At present, such solutions are only developed by consulting firms that would involve more expenditure and a difficult case for the value-for-money argument.

The main aim of this project is to leverage artificial intelligence and machine learning to develop a text analytics service that can deliver sentiment analysis and predictive topic modelling from the social media customer interactions through a cloud service. The results from the service would then be used to inform the decision-making process by providing output that will be informative. The prototype solution will be in line with the current strategic plan of delivery online services with new technologies and helping to achieve the goal of being in the upper quartile in customer service and experience.

The project will collect Twitter data addressed to the organisation and use the data to develop a machine learning model. The model will then be deployed as a cloud webservice and consumed through a website that can then be used by the customer service teams to analyse social media posts from the Twitter and Facebook accounts.

A review of the twitter data found that it is possible to develop a predictive topic model that narrows down the topics to 8 generalised topics. The sentiment of the customer interactions is negative which is expected as customer service is an avenue of handling issues when they go wrong.

In conclusion, it is feasible to develop a cloud service that provides sentiment analysis and predictive topic identification in the current cloud environment. It will also contribute positively in assisting to achieve the goal of achieving upper quartile in customer service by providing real-time results of the sentiment analysis by topic. The delivery of such a service will also be a sustainable solution by leveraging the environmental benefits of cloud computing.

Contents

Declaration.....	i
Acknowledgements.....	ii
Abstract.....	iii
Contents	v
List of Figures.....	viii
List of Tables	x
Glossary of terms.....	xi
1 Introduction	1
1.1 INTRODUCTION.....	1
1.2 BACKGROUND	3
1.3 AIMS AND OBJECTIVES:	4
1.3.1 <i>Aims</i>	4
1.3.2 <i>Objectives</i>	4
1.4 FOCUS	5
1.5 STRUCTURE	6
2 Literature Review.....	8
2.1 ARTIFICIAL INTELLIGENCE.....	8
2.2 MACHINE LEARNING.....	11
2.2.1 <i>Supervised Learning</i>	13
2.2.2 <i>Unsupervised Learning</i>	14
2.2.3 <i>Semi-supervised Learning</i>	14
2.2.4 <i>Reinforcement Learning</i>	15
2.2.5 <i>Bias in Machine Learning</i>	16
2.3 TEXT ANALYTICS	17
2.4 NATURAL LANGUAGE PROCESSING	18
2.4.1 <i>Sentiment Analysis</i>	18
2.4.2 <i>Topic modelling</i>	20
2.5 DATA.....	22
2.5.1 <i>Structured data</i>	22
2.5.2 <i>Unstructured data</i>	22
2.6 CLOUD COMPUTING	23
2.6.1 <i>Platform as a Service (PaaS)</i>	24
2.7 DATASETS	26
2.8 DATA SCIENCE	26
2.8.1 <i>Software</i>	27
2.9 CONCLUSION	29
3 Research Methodology.....	31
3.1 INTRODUCTION.....	31
3.1.1 <i>Philosophy</i>	31
3.1.2 <i>Approaches</i>	32
3.1.3 <i>Strategy</i>	32
3.1.4 <i>Risks & Limitations</i>	33
3.1.5 <i>Validity, Reliability & Generalisability</i>	33

3.1.6	<i>Data Evaluation Methods</i>	34
3.2	DEVELOPMENT METHODOLOGY	35
3.2.1	<i>Software Development Methodology</i>	35
3.3	RESOURCES	35
3.3.1	<i>Hardware Resources</i>	35
3.3.2	<i>Software Resources</i>	35
3.3.3	<i>Human Resources</i>	37
3.4	CONCLUSION	37
4	Software Requirements Specification (SRS).....	38
4.1	INTRODUCTION.....	38
4.2	GOALS OF SRS:.....	38
4.3	FUNDAMENTALS	38
4.3.1	<i>Functional Requirements:</i>	38
4.3.2	<i>Non-functional Requirements:</i>	39
4.3.3	<i>Use-cases:</i>	39
4.4	PROCESS.....	40
4.5	VALIDATION.....	41
5	Design and Development	42
5.1	DESIGN.....	42
5.1.1	<i>Design Outlook</i>	43
5.1.2	<i>Data Collection</i>	44
5.1.3	<i>Data Pre-processing</i>	45
5.1.4	<i>Stop Words</i>	46
5.1.5	<i>Extracting features</i>	48
5.1.6	<i>Data Visualisation</i>	48
5.1.7	<i>Modelling</i>	48
5.2	DEVELOPMENT.....	50
5.2.1	<i>ML model</i>	50
5.2.2	<i>Production API</i>	60
5.2.3	<i>Database Development</i>	62
5.2.4	<i>UI Development</i>	63
5.2.5	<i>Dashboard development</i>	63
6	Testing	65
6.1	MANUAL TESTING.....	65
6.1.1	<i>Test 1:</i>	65
6.1.2	<i>Test 2:</i>	66
6.1.3	<i>Test 3:</i>	66
6.1.4	<i>Test 4:</i>	67
6.1.5	<i>Test 5:</i>	68
6.1.6	<i>Test 6:</i>	68
6.1.7	<i>Test 7:</i>	69
6.1.8	<i>Test 8</i>	70
6.1.9	<i>Test 9</i>	70
6.1.10	<i>Test 10</i>	71
6.1.11	<i>Test 11</i>	71
6.2	CONCLUSION:.....	72
7	Results and Evaluation	73
7.1	TEST RESULTS.....	73

7.2	SENTIMENT SCORES.....	75
7.3	TOPIC MODELLING	77
7.4	GROWTH OF SOCIAL MEDIA COMMUNICATION	77
7.5	CURRENT STATE.....	79
8	Discussion	84
9	Conclusion.....	86
10	Reflection.....	87
11	References	89
12	Appendix	99
12.1	SRS.....	99
12.1.1	Use Cases:.....	99
12.2	DESIGN AND DEVELOPMENT	99
12.2.1	Pre-processing.....	99
12.3	DEVELOPMENT.....	105
12.3.1	AWS Lambdas – based in Ohio	105
12.4	RESULTS.....	108
12.4.1	Sample Text for testing	108
12.5	FRONTEND LINK.....	111
12.6	AZUREMLAAS LINK	111
12.6.1	Swagger API info.....	114

List of Figures

Figure 1 Sentiment Analysis techniques [77]	19
Figure 2 Graphical model representation of LDA [81]	21
Figure 3 The Shared Responsibility Model [90]	24
Figure 4 The Data Science Venn Diagram [104]	27
Figure 5 Sequence diagram of the prototype with the analyst actor	40
Figure 6 Use of prediction and optimisation to add value in business systems [111]	42
Figure 7 Prototype system design	43
Figure 8 Original Ubuntu dataset.....	44
Figure 9 Example of TweetScraper tweet file	44
Figure 10 Example of importing data	46
Figure 11 Example of normalisation	46
Figure 12 Example of tokenization	46
Figure 13 Example of stop words removal	47
Figure 14 Stemming Techniques [115]	47
Figure 15 Example of stemming	48
Figure 16 Example of extracting parts of speech features	48
Figure 17 Modelling overview [118]	49
Figure 18 Example of using Vader in NLTK	49
Figure 19 Iteration 1 results	51
Figure 20 Iteration 2 NMF results.....	51
Figure 21 Iteration 3 Sentiment Analysis.....	52
Figure 22 NMF and LDA modules.....	52
Figure 23 Results of running NMF and LDA.....	52
Figure 24 LDA as normal	53
Figure 25 LDA after using LDA Mallet	53
Figure 26 Topics from iteration 5	54
Figure 27 Topic modelling results	55
Figure 28 Azure dataset upload	56
Figure 29 Machine learning development tasks	56
Figure 30 AMLS Pre-processing	57
Figure 31 LDA and K-means options in AMLS	57
Figure 32 AMLS final training experiment	58
Figure 33 AMLS predictive service.....	59
Figure 34 Testing webservice endpoint with Postman	59
Figure 35 Deployed Service.....	60
Figure 36 Sentiment lambda	61
Figure 37 Azure lambda.....	61
Figure 38 PostgreSQL lambda	61
Figure 39 Amazon API Gateway.....	62
Figure 40 UI frontend of prototype.....	63
Figure 41 Quicksight Workflow [126]	64
Figure 42 Connect PostgreSQL to Quicksight.....	64

<i>Figure 43 : Example Twitter Text "tweet"</i>	68
<i>Figure 44 Comparison of sentiment scores</i>	76
<i>Figure 45 Comparison of predictive topics</i>	77
<i>Figure 46 Growth of Twitter Communications relevant to DVLA</i>	78
<i>Figure 47 Average percentage probability per topic per year</i>	79
<i>Figure 48 Topic 3 vs Sentiment</i>	80
<i>Figure 49 Topic 4 vs Sentiment</i>	80
<i>Figure 50 Topic 6 vs Sentiment</i>	81
<i>Figure 51 Topic 8 vs Sentiment</i>	82
<i>Figure 52 Topic 2 vs Sentiment</i>	83
<i>Figure 53 LDA Topic 1</i>	101
<i>Figure 54 LDA Topic 2</i>	102
<i>Figure 55 LDA Topic 3</i>	102
<i>Figure 56 LDA Topic 4</i>	103
<i>Figure 57 LDA Topic 5</i>	103
<i>Figure 58 LDA Topic 6</i>	104
<i>Figure 59 LDA Topic 7</i>	104
<i>Figure 60 LDA Topic 8</i>	105

List of Tables

<i>Table 1 NLTK modules for language processing tasks [108]</i>	29
<i>Table 2 Example of Functional requirement</i>	38
<i>Table 3 Example of Non-functional requirement</i>	39
<i>Table 4 Example of a use case</i>	39
<i>Table 5 Actors of the prototype</i>	40
<i>Table 6 Relevant word(s) in the predictive topics</i>	73
<i>Table 7 Test coverage</i>	74
<i>Table 8 Comparison of Sentiment Results from the system and expert opinion</i>	75

Glossary of terms

AI – Artificial Intelligence

API – Application Programming Interface

BI – Business Information

CRM – Customer Relationship Management

DBaaS – DataBase as a Service

DBN – Deep Belief Networks

DVLA – Drivers and Vehicles Licencing Agency

EM – Expectation Maximization

GMM – Gaussian Mixture Model

IaaS – Infrastructure as a Service

JSON – JavaScript Object Notation

LDA – Latent Dirichlet Allocation

MDP – Markov Decision Processes

MI – Management Information

ML – Machine Learning

MLaaS – Machine Learning as a Service

NLP – Natural Language Processing

OS – Operating System

PaaS – Platform as a Service

PCA – Principal Component Analysis

REST - Representational State Transfer

SaaS – Software as a Service

SVM – Support Vector Machines

1 Introduction

In this project, we will consider how an organisation can make effective use of the available unstructured data such as social media customer interactions to better understand its customers by using artificial intelligence and machine learning. The solution will involve a prototype development of a machine learning model deployed for sentiment analysis and topic modelling delivered using cloud services technology.

This chapter will provide an introduction to the issues faced by organisations when it comes to unstructured data (1.1), provide a background to the problem domain as experienced (1.2), define the specific focus issue (1.3), and provide a structure to the report (1.4).

1.1 Introduction

Data is at the heart of every organisation. Without data, then the business of an organisation is pointless and rudderless. Every organisation from the smallest business of one to the largest multinational businesses of the world are dependent on data and better still on the quality of it. But having the data is only the start of the journey to be a successful business. It is the part of knowing what to do with the data through analysis that provides the greatest advantage and provides the differentiating factor in the success of similar businesses.

For years, organisations have been collecting data and running analysis against it. This has been made possible by making use of data that is clearly defined in the description and saved in systems for later use and interrogation. This data is known as structured data. But structured data that can easily be defined and saved in systems only cover part of the available data that any organisation collects as it generates both as a consumer and a

generator. Since the exponential growth in the use of the Internet, the amount of data generated has mirrored that growth. Since the early 2000s, it has been known in the industry and reported through industry journals [1], research [2] and umpteen reports [3] that unstructured data was becoming a problem for organisations. It became evident that organisations were only dealing with a small part of the data relevant to them and were being overwhelmed by how to handle the majority of the data they were in contact with. Surveys have found guesstimates of around 80% [3] in usual business data and 95% of big data¹ [4] to be of the unstructured form. But what exactly is unstructured data?

Unstructured data is any data that fails to fulfil a standardised specification. Unstructured data can take the form of emails, support tickets, chats, social media post, call centre logs, call centre notes, videos and audio communication bulletins and old fashioned business letters [5]. The nature of this data makes it difficult to interrogate as it is inherently unstructured, with no obvious headings, descriptions and in some cases without an obvious saved form. The current favoured form that gets saved in a database is the structured data. This allows an organisation to interrogate this data through the well-established methods of queries. This method of querying cannot be translated easily to the unstructured data. Blumberg and Atre [1], found that a lot of unstructured data happen to exist in customer relationship management(CRM) systems and particularly in call centre staff interactions with customer records. It is a similar experience in medical records [6] where unstructured data is prevalently produced and contains critical information.

We will highlight the current methods being employed to infer meaning from unstructured data. The project will show how much meaning can be inferred from unstructured data by making use of natural language processing tools through machine learning models. The implementation will look at two forms of unstructured data i.e.

¹ Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation. [136]

support chat logs and social media posts. The experience gained when developing using support chat logs will then be used to iterate to a prototype system that will make use of social media posts. The use of social media text will provide a varied picture of what consumers are experiencing while meeting the need for catering for a system to cope with the growth of social media as a means of communication [7]. The methods employed will include information extraction through the use of feature extraction, sentiment analysis and topic modelling.

1.2 Background

As highlighted, organisations are still yet to come to terms with the age of big data. In this new age, customers are interacting with an organisation in ways that are most convenient to them. Research shows that more customers are choosing to interact with the business through the use of social media [8] rather than the traditional methods for example via letter through the post. This is proving to be a challenge as although organisations have provided a presence in the online world through accounts in popular social media sites or even host online contact avenues, the old methods of working and interactions for example that involving capturing customer details into customer relationship systems and tools is still heavily relied upon and maintained. However, this allows for opportunities for the true “voice of the customer” and context to be missed. In this new age, it has meant that the organisation needs to be aware of the “voice of the customer” as this “voice” can be easily be shared with rest of the world involving other customers, competitors and rivals with unmitigated context.

It is therefore in the interest of the organisation to develop new systems that can harness this new way of customer interaction. It is such systems that text analytics has become popular with organisations. According to the Forrester [9], demand for text analytics has increased in line with the increase of unstructured data held by organisations. The current development in the field of text analytics is to deploy enterprise solutions that can be used in real-time or can be batch streamed to run the analysis. The solutions available are

however expensive and includes a hidden technical debt [10] that overshadows the deployment of such systems.

1.3 Aims and Objectives:

1.3.1 Aims

To investigate how, why and where the integration of machine learning and artificial intelligence in customer management gives business value by being able to find and demonstrate the sentiment from social media interactions.

To investigate how, why and where the integration of machine learning and artificial intelligence in customer management systems gives business value by providing derived topics from customer social media interactions.

1.3.2 Objectives

The objectives of this project are to:

1. produce a machine learning model capable of undertaking text analytics of sentiment analysis and predictive topic identification on the submitted social media text
2. ensure the model can be deployed in the cloud and provide RESTful APIs for integration
3. make use of cloud technologies such as serverless in the deployed system
4. develop a frontend to provide an interface to the deployed text analytics web service
5. provide a management information (MI) and or business information (BI) dashboard to review the deployed current system status of the sentiment by topic.

1.4 Focus

We will consider the case of a public organisation, DVLA [11], that deals with citizens by providing services of registering and licensing of drivers and vehicles and the collection of vehicle excise duty on behalf of the UK Treasury. The agency according to the strategic plan for 2017-2020 sets out an ambition of developing a hub for digital motoring through the use of more agile and cloud services. The plan highlights the current take-up of digital services at 92.3% which highlights the extent of the Agency's interaction with its customers digitally. A major goal of the plan is to provide best-in-class customer services. The question then arises, how best can the Agency get to leverage the unstructured data currently being created during its legal service?

Text analytics offers an answer to this question. The field of text analytics has exponentially grown with more organisations adopting text analytics to identify and gain insight from textual information. Gartner [12] reports that text analytics has been heavily adopted by industries such as banking [13], healthcare and retail. M&M [14] estimates that the text analytics market will grow from \$3.2 billion in 2016 to \$8.8 billion in 2022, more than double the spend during the period. One of the major market reasons for this growth has been the growth in the need for predictive analytics. According to one of the learning analytics providers SAS, predictive analytics is the use of data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data [15]. In predictive analytics, organisations use models to derive information such as sentiment that informs on hidden trends and dependencies.

It is in this same area that this project will seek to show that it is possible to provide such a service that will also be in line with market trends as a cloud-based text analytics service. The use of data visualisation will also be part of the prototype and will show it as a relevant part of the service as it best provides the insight in an easy-to-consume format for management information.

The Agency in the recent past has developed new services that have included making use of social media and new call centre services. As part of the current call centre system, a voice analytics system is in place with planned future development on text analytics yet to be deployed. However, that analytics does not cover elements outside the voice call framework.

The context, therefore, is to develop a text analytics system that could be used to identify the sentiment of the customer's communications i.e. social media text and then categorised it into topics.

The iterative development will involve developing two systems by starting using the chat support logs and then make use of the experience through refactoring the ideas into the prototype system using Twitter² data. The dataset used for the chat logs is one that is publicly available as a data corpus from the Ubuntu Dialogue Corpus [16]. For the social media posts, the project purposely created a dataset from Twitter tweets based on the publicly available @DVLGOVUK handle and #DVLA hashtag interactions from the dates of 05-01-2009 to 21-09-2019.

1.5 Structure

The report will be laid out in different chapters starting with a background literature review chapter looking at artificial intelligence, machine learning, text analytics, data and natural language processing (in particular on sentiment analysis and define why it is important to text analytics), cloud computing, data science and datasets. There will then follow chapters on project planning, and research methodology which will describe the planning and methods used in the report. A chapter on software requirements will explain the requirements of the system. A chapter on design and development will describe the software development process including the choices made on process and software

² Twitter – is an internet microblogging and social networking service that allows registered users to send messages known as “tweets” or publish a post on their profile.

choice. A chapter will then follow showing the tests done on the developed system and this will then lead on to a chapter to evaluate the results, expectations and highlight the issues encountered. In conclusion, we will provide a summary of the results, code, weblink of the prototype developed and offer recommendations on the process, system and experience gained and discuss further future work. An appendix will be included which will include code, datasets and all associated documents used in the report.

2 Literature Review

2.1 Artificial Intelligence

“AI began with an ancient wish to forge the gods” [17]

Intelligence is inherent to the description of who and what mankind is. Without intelligence, it is difficult to appreciate mankind as being different from other species. What makes intelligence has therefore been an intriguing question. How are humans able to record their thoughts? The uniqueness of mankind is that the ideas of philosophy, mathematics, understanding of logic, psychology and linguistic get raised, studied, debated and agreed. The emergence of machines has provided tools for evaluating these ideas and therefore making artificial intelligence materialise.

Artificial intelligence, therefore, owes its growth and prevalence to the rise of machines that are now referred to as computers. The term artificial intelligence was first used in 1956 by Professor John McCarthy during the Dartmouth conference [18]. During the Dartmouth conference, the idea of artificial intelligence as we currently understand was conceptualised. Professor McCarthy defined artificial intelligence (AI) as

“ ... the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but artificial intelligence (AI) does not have to confine itself to methods that are biologically observable.”

Another definition of artificial intelligence is that it is the field of study that looks at the techniques for undertaking tasks central to human intelligence [19]. Human intelligence is therefore fused with the biological description of who and what a human being is. Artificial intelligence can then be argued to require the understanding of the thought process, performance and outcome achieved by a non-human thing. The non-human thing would simply be defined as a machine, computer or expert system that accepts some

inputs. For any intelligent expert system to be measured, the measurement would have to be as a measure of the outcomes of performance. Therefore, the defining question that needs to be answered has to be “how did the system get the task right given the inputs?”

Russel and Norvig [20] summarised the categories of artificial intelligence into four i.e. thinking humanly, acting humanly, thinking rationally and acting rationally. The categories of an AI system, therefore, require the use of a hypothesis and experiments to prove it. The rational categories such as mathematics [21] [22] and engineering [23] [24] proved to be the earlier categories of advancing the ideas of AI. It is during a difficult period at the height of the Second World War that Turing [25] discussed conditions that would make a machine intelligent. Turing’s interest was how such a machine would need to act to meet the threshold to be considered as acting humanly. In the now-famous test of the imitation game, later named after Turing, Turing proposed that for a machine to be considered intelligent it needs to pass a test by answering questions posed to it by a human. For such a machine to pass the test it needed skills that included natural language processing, knowledge presentation, automated reasoning, and machine learning. The test was further adapted and extended to include aspects of vision and movement. The latter being known as the total Turing test.

Another aspect of ability is the ability to be creative. Boden [26] critiqued the available developed AI systems in their ability to produce creativity. Creativity is measured by the production of ideas that are considered new. The idea of “new” can be considered from the viewpoint of within the system or from outside the system. This is what Boden argues as P-creativity and H-creativity and where P stands for psychological and H for historical. AI models have primarily been set up to concentrate on delivering P-creativity with H-creativity a happy coincidence outcome of most systems. Creativity from these systems results in the development of familiar, exploratory and transformational ideas. Boden [26] identifies that for any of these ideas to be generated the systems are dependent on domain expertise and analytical power. Domain expertise has been shown to advance

development [27] [28] in as much as analytical power development coupled with technological improvements [29] has led to the development to better models.

Development in AI, mainly through research, has focussed and morphed into many fields chiefly focussing on intelligence in learning, problem-solving, reasoning, natural language, vision and perception. Some of these same areas were identified in the early 1960s as areas of problems and opportunity. Minsky [30] identified in his paper those areas that prevented a computer to be considered as intelligent were primarily around search, pattern-recognition, learning, planning and induction. The areas that form part of artificial intelligence include machine learning, reasoning, planning and natural language processing.

Machine learning is the study of data-driven methods capable of “learning something useful” through tasks from basic to complex using systems that require some learning [31]. The earliest artificial intelligence systems in the world involved simple computer programs that required learning through a method of trial and error like the Turing machine [25], Minsky machine [32], and the General Problem Solver machine [33] [34]. We will further discuss machine learning in Section 2.2.

Reasoning involves the use of logic and knowledge in the decision-making process and uses deduction and induction in the process. These systems are used in an agent and in multi-agent system models where the importance of knowledge and common knowledge is paramount to ensure success [35] .

Planning is about decision-making processes that involve the determination of a solution for a declared problem while ensuring that the performance of the system will be the “best” option that provides the best outcome [20]. The determination of “best” might be in terms of processing power, time, speed, accuracy and other determined significant measures. Planning problems are particularly prevalent in manufacturing industries and

areas that lend themselves to automation like risk management, cybersecurity and transport planning. One of the earliest planning systems was the “Shakey the Robot” [36].

Natural language processing (NLP) is the field that considers machine comprehension and manipulation of human language. Language is that uniquely human trait that is used to provide instruction. Language in AI focuses on the use of words as input. Such systems are said to be involved in natural language processing. The earliest systems in NLP were used in the Cold War years in attempts to automatically translate language through mechanical translations [37]. The systems were largely inaccurate and required processes of pre- and post-processing to get any usable accuracy of the texts. Later systems such as Eliza [38], and SHRDLU [39] were developed that improved the ability of the systems to “accept” the language using the rules of the language. However, systems that provide a complete human-like language capability have been difficult to develop [20]. Recently major technology companies like Apple with Siri, Amazon with Alexa and Google with Google Assistant are currently working to improve the use of a variety of voice assistants that make use of NLP methods such as identifying an entity, named entity and entity extraction but are still far from perfect [40].

2.2 Machine Learning

While AI is the science of trying to copy mankind’s capabilities, machine learning (ML) is the science of training a system through the process of learning over time from available data. The ultimate aim of the training is to improve the outcomes of the system. Samuel [41] defined the term machine learning as the field of study that gives computers the ability to learn without being explicitly programmed. Mitchell [42] proposed a more scientific definition for machine learning as:

“... a computer program is said to learn from experience E with respect to some task T and some performance measure P , if its performance on T , as measured by P , improves with experience E .”

Another definition, Murphy [43] offers it as:

“... a set of methods that can automatically detect patterns in data and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty such as planning how to collect more data!”

The iteration of tasks is necessary to finesse the responses from the models. The steps require making use of previous response to inform on the next steps required. After several iterations, the developing model can give a better-accepted response. Deciding on the steps in training the model leads to the question of how those steps affect the training. What type of learning achieves the best outcome? To carry out the learning, algorithms have been developed to be effective depending on the desired outcome.

An algorithm is a program that takes an input and gives an output dependent on some defined logic [20]. The algorithm is defined as a set of instructions to define how to work the data, however, in machine learning the process of interaction is counter-intuitive in that the data defines the final instructions (model). The machine learning categories based on algorithms (i.e. dependent on the outcome) are mainly regression, classification or clustering algorithms.

Regression algorithms include linear regression, Bayesian linear regression, neural network regression, and decision tree regression [44]. Classification is either two-class or multiclass. For two-class classification algorithms include two-class support vector machine (SVM), two-class logistic regression, and Naïve Bayes classifier. For predicting three or more categories the algorithms include one-v-all multiclass, multi-class decision forest, and multiclass neural network among others. Clustering algorithms are algorithms that have been developed for grouping points that prove to very efficient in identifying the underlying structure of a population. They include k-means [45] and the variant k-means Bradley-Fayyad-Reina [46], Expectation-Maximization(EM) using Gaussian Mixture Models (GMM) and Principal Component Analysis(PCA) [47] [48] among others.

It is also important to define machine learning depending, on the human input and interaction with the systems. Such criteria would provide categories such as supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning.

Another criterion would be whether the system received learning input from an online feed [49] or updated through a batch learning system [50]. The use of patterns to develop models resulting in a model-based learning system [51] or an instance-based system [52] can also be used to identify the ML systems.

2.2.1 Supervised Learning

Supervised learning is the machine learning task where the data is split between training and testing data where the training data is labelled to identify the data according to a category or label giving it a pair of input identifier and the desired output identifier [20]. Labels will, therefore, be the desired outcome of the model output. The labels provide the model with what to learn by example. The system is also said to be learning by induction [53].

In the data, it is vital to identify the attributes and features. An attribute in machine learning is the data type, while the feature is more than the data type and includes more information of a measure of value [54] such as are the values dependent or not. If the feature is used to predict a value, then that feature is called a predictor.

Classification is used where the data can produce discrete response e.g house prices or pay levels. Regression is used where the dependent variable is not discrete e.g. male or female, mango or orange. The common algorithms available for use in supervised learning include linear regression, Naïve Bayes [55], logistic regression, k-Nearest Neighbours, and support vector machines (SVMs) among others.

2.2.2 Unsupervised Learning

Unsupervised learning is the method of learning where the training data used contains no labels and it is left to the model to identify features as it develops [54]. Algorithms used in unsupervised learning include clustering algorithms like k-means and expectation-maximization, visualisation and dimensional algorithms like principal component analysis (PCA). Unsupervised algorithms work well where it can be able to find a commonality in the data. The algorithms work by sub-dividing the data (hierarchical clustering), simplifying the data (dimensionality reduction), and identifying anomalies by learning associations.

k-means is one of the simplest clustering methods that is effective as an unsupervised learning algorithm [54]. The number of clusters, k , is chosen in advance and then the task of the model is to assign inputs to the cluster that has the least mean distance to the centre.

2.2.3 Semi-supervised Learning

This is the case where some of the training data involved contains partially labelled training data. Algorithms used in semi-supervised learning involve a combination of unsupervised and supervised algorithms [54]. An example of such a system makes use of stacked neural networks known as deep belief networks (DBN) running in stacks of machines known as restricted Boltzmann machines (RBMs). A neural network is a group of algorithms that mimics the human brain that works together for a purpose. The network in its simplest form is made up of a visible layer, a hidden layer, output layer and nodes. A node is a point where the computations happen after some activation, and where the operations mimic the workings of the neuron in the brain.

The advantages of using semi-supervised learning are that it allows for the use of unlabelled and labelled data together in one label. Chapelle, Scholkopf and Zien [56] explained the required assumptions and the rules for semi-supervised learning to work. It

relies on some assumptions to be present for the learning to work and happens to be dependent on the input. The semi-supervised smoothness assumption is where the output of corresponding outputs (y_1, y_2) expects the inputs (x_1, x_2) to be close. The cluster assumption where if the points in the data are in the same cluster then they are likely to be in the same class in the data. The cluster assumption is however dependent on the low-density separation rule that defines the decision boundary to define the clusters, where the boundary is defined that it should lie in a low-density region. The manifold assumption states that data falls on a low dimensional manifold. Data falling on such a manifold avoids on the curse of dimensionality [57], which relates to the exponential growth of dimensions increasing the volume making the data sparse affecting significance of the data points, by reducing the dimensionality [58]. Semi-supervised learning works best in situations where obtaining a large data set is easy but labelling the data set would involve a huge effort. Such data sets include web page classification, speech recognition data and genome sequencing projects.

2.2.4 Reinforcement Learning

Sutton and Barto [59] define reinforcement learning as goal-directed learning achieved through interaction. The system is not directed to take actions but instead receives feedback to any actions taken and given a reward. In time the system discovers what actions leads to the most beneficial reward. To understand how reinforcement learning works, the system has a learning agent, the system environment, value function, and the goals that determine the size of the reward. The learning agent is tasked with achieving the goal while getting to understand the environment through the feedback received over time from outside the environment with regards to the actions taken to get to the goal. The value function defines whether over time the actions taken have received a maximum reward. Reinforcement learning makes use of Markov Decision Processes (MDPs) [60] [61] which define the working between the learning agent, environment, reward and actions to get the rewards. MDP is a random process where the probability of the result of actions is not dependent on past states and results.

Reinforcement learning is different from supervised learning because in supervised learning the system is provided with training label examples at the start. Reinforcement learning is also different from unsupervised learning because the system receives feedback as it learns to generalise the solution to the goal. The major challenge of reinforcement learning is finding the right balance between letting the system explore and exploit. The balance challenge is getting to a situation where the learning agent can retake previous learning actions that have resulted in reward and also try new actions that provide the possibility of a better reward.

A recent development that has shown the success of reinforcement learning is AlphaGo Zero [62], which is an improved version of AlphaGo. AlphaGo has the distinguished title of being the first program to defeat a human world champion in the game of Go. It is the first successful self-trained algorithm that learned on a dataset that had not been curated by human experts. It showed the art of possible by removing the need for human-curated learning datasets.

2.2.5 Bias in Machine Learning

The biggest issue in machine learning models is the existence of bias. Bias is an intrinsic component of the makeup of the data and the workings of machine learning. Hildebrandt [63] argues that the best way to mitigate the issues of bias, either in data or design, is for a research design methodology that involves the users and developers of machine learning models to learn more about the potential and limitations of the machine outcomes. It is therefore vital that issues of bias are addressed as more models start to be deployed in society especially in areas of law [64] and medicine. How to resolve the issue of bias, therefore, becomes an integral part where the selection and identification of the data including the determination of the interest in the data, the choice of the algorithm to use and the training management in the steps of producing the model are carefully considered.

It is with this in mind that a lot of effort by machine learning researchers [65] and experts are focused on highlighting the problems associated with bias and developing tools to mitigate it. Dixon et.al [66] looked at unintended bias in text and found that one way of mitigating bias in text classification is counter-balancing the effect by adding data from other domain sources. Tools such as the IBM AI Fairness 360 toolkit [67], FairML [68] and others are being developed to assist in the identification of bias and sharing best practise to tackle bias.

2.3 Text Analytics

Text mining is the process of refining data collected over some time into a usable format using tools [69]. This involves processes of taking the initial form of the data and through steps of pre-processing, text mining operations and postprocessing to be able to derive value. Hearst [70] defines text mining as,

“ ... the discovery by computer of new, previously unknown information, by automatically extracting information from written sources”.

Text analytics is normally referred to as text mining or machine learning from text depending on the context. Text mining differs from information retrieval as rather than just getting the source of the information (documents or texts), it further adds useful data, known as metadata, about the documents (texts). Text mining is best carried out by software that can decipher the metadata in the texts and able to provide visual aids of the information [71] [72].

Text analytics is the extraction of useful insights by the application of statistical tools and machine learning methods and algorithms to discover any discerning features in the text [73]. In the Focus section in the Introduction, we highlighted the growth in the field of text analytics. It has exponentially grown with more organisations seeking to adopt text analytics to effectively make use of its data. Industries that have heavily adopted text

analytics include banking, healthcare and retail [12]. Future prediction indicates that the use of text analytics will grow, and form part of core business spend [14].

2.4 Natural Language Processing

Liddy [74] offers the following definition for natural language processing (NLP),

“... is a theoretically motivated range of computational techniques for analysing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.”

NLP, therefore, includes in part natural language understanding and natural language generation. Natural language understanding is taking an input of a text and being able to make an output that is useful and provides true representation. A good example of this is the translation of a text from dialogue in a movie from the spoken language to the user preferred language through the use of subtitles. Natural language generation is the task of generating text or speech in a language from resources held in machine format. Generation of language in this way has however invited criticism and concern as this process bypasses human input, and language is considered an intrinsic part of being human. An example of a natural language generator is the Washington Post system, Heliograf [75], used to write new reports during the 2016 Summer Olympics and in the 2016 United States Election. In the 2019 General Election, the British Broadcasting Corporation (BBC) used a semi-automated system, Semi-Automated Local Content (SALCO), to auto-generate election result news reports [76].

2.4.1 Sentiment Analysis

Sentiment analysis (SA) is also known as opinion mining (OM). Sentiment Analysis is the computational study of people’s opinions, attitudes and emotions toward an entity

[77]. Therefore, the importance of undertaking SA is to find opinions, thoughts, views or expressed sentiment. Organisations undertake SA on their customer or product or service reviews or recommendations to find out how well or not their service or product is truly performing. SA is considered a classification task where it is important to undertake specific steps to finding the sentiment. To undertake a SA task, features from the text are extracted. Some of the features extracted from the text include negation which easily identifies negativity. Or parts-of-speech (POS), where adjectives are used to indicate opinions [78].

Figure 1, shows the techniques of sentiment analysis. Sentiment analysis can either be done through using machine learning or through lexicon approach. Lexicon based approach requires the creation of a lexicon dictionary with the relevant sentiment polarity [78]. The use of machine learning algorithms to undertake SA and document analysis has been extensively researched with a particular interest in supervised algorithms [55] [79]. Unsupervised algorithms have also been used in developing models for sentiment analysis, especially where labelling in the corpus does not exist. Models such as one done using Latent Dirichlet Allocation (LDA) [80] has been shown to work in unsupervised machine learning situations. LDA is further discussed in the following section as part of topic modelling.

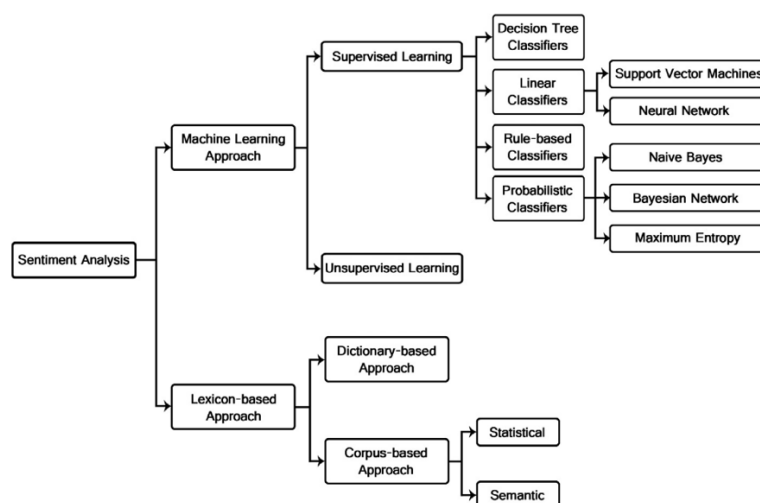


Figure 1 Sentiment Analysis techniques [77]

2.4.2 Topic modelling

Unstructured data as highlighted in the introduction is the bane of organisations. This is especially true for those situations where the opinion of customers and consumers is sought. It is usually the aim of such an exercise to further glean meaning at best to simplify topics that can provide a simplistic view. The predominant technique and methodology for undertaking this exercise is topic modelling.

Topic modelling is about taking a text corpus and uncovering topics and discovering latent hidden themes in the text. The task is also about being able to provide an insight that can point to the sentiment and possibly to the emotion of the text. The nature of natural language text is the fact that the subject, context and phrasing can hugely affect the reading of the text. Topic modelling is therefore widely deployed in an organisation that relies on being informed of customer feedback. Topic models are computer programs that statistically take the existence of a series of words and derive a topic in the view of the computer. It is important to recognise that the association of words to the topic can vary from the view of a human being and the computer. The success of a modelling task is dependent on a number of things, including the quality of text processing, having a variety of topics to talk about, choice of an appropriate modelling algorithm, deciding on the number of topics to ask the algorithm, and deciding the tuning parameters.

One example of a topic modelling method that has been widely used is the Latent Dirichlet Allocation (LDA), which was highlighted as one of the methods also used for sentiment analysis in the previous section. LDA [81] is a generative probabilistic Bayesian model for use with the collection of discrete data such as text corpora. Generative means that the model produces predictive data using a joint probability function of the data input values. The approach of LDA is summarised by its two core principles central to its methodology.

The first principle is that for every document (text) there is a mix of random latent topics. To illustrate, take a statement such as “gaming has been increasing over the years in line with internet use and some people comment that this has resulted in the rise of crime”. That text has three topics: gaming, internet and crime. The number of topics and the influence of the topics to the overall reading of the text varies with each text. The second principle is that every topic is formed from a body of words. An example of this would be for example a topic television would include words about “shows”, “series”, “news”, and “streaming”. LDA is, therefore, a mathematics methodology of estimating both the mix of words in a topic and the mix of topics in a text corpus. A review [82] of various published work about the implementation and usage of LDA through modification show that LDA through its resilience and popularity in the field of text analytics is a cornerstone method in topic modelling.

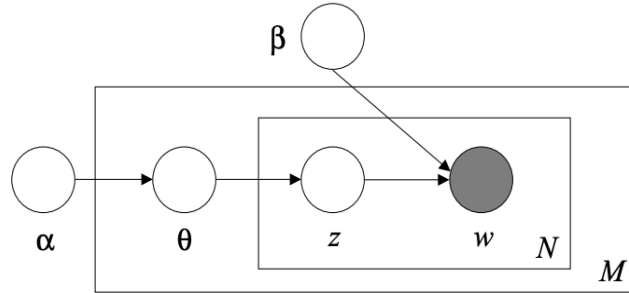


Figure 2 Graphical model representation of LDA [81]

The parameters α and β in **Figure 2** are text-level corpora parameters that are sampled once in the generation of the corpus. The θ is overall text variables that are sampled once per document while the z and w are word-level variables (topics z , words w) sampled once for each word in each text document where the two plates N and M are representative of replicates with M representing the overall text (document) and the inner plate N represents repetition in topics and words present within [81].

2.5 Data

Data, as we mentioned in the Introduction, is at the core of all organisation, public or private. Data is the raw collection of information that is collected in its impure state that forms the basis of an organisation process. Data can be categorised as either structured or unstructured or a hybrid of the two.

2.5.1 Structured data

Structured data is easily organised information that can be held in a form that can be interrogated easily. Structured data has the property of being able to be query and stored in a format that is easy to enter, store, query, analyse and visualise. Structured data is usually saved in a database and managed through the use of structured query language (SQL) that was specifically developed to handle structured relational database systems.

2.5.2 Unstructured data

Davis [3] in a recent opinion lead for CIO says that at least 80% - 90% of all business data is unstructured data. Daily statistics show that 500 million tweets, 294 billion emails, 4 petabytes on Facebook, 4 terabytes on transport links, 28 billion petabytes from wearables, 65 billion messages on WhatsApp and 5 billion searches on search engines [83] are the current state of measurable data. It shows that most of today's data held as pictures, documents, video, web data such as social media messages and audio files are held as unstructured data.

The task for any organisation then is to ensure that it can organise the available daily data glut. This huge collection of data is what has become known as big data. The recent digital explosion has meant the “day-to-day” business is no longer just a part of the day business i.e. 9-5 but an “always-on” 24/7 business. The problem as Davis highlights is not that organisations are unaware of the amount of data but rather that the tools and technologies

available to make sense of the data are lacking. Organisations are however taking steps to try and mine the unstructured data for useful information to inform on their decision making and improve their competitiveness. At present this is proving to be an impossible task to undertake without the help of service companies as highlighted in the introduction. Adnan et.al [84] highlighted the experience of healthcare in making sense of the unprecedented growth in unstructured data. An IDC report [85] shows that the trend in the coming future is for an increase in organisational data meaning that organisations that want to survive the data tsunami need to have a solution to handle the unstructured data.

Organisations can take advantage of AI, ML and using tools such as topic modelling and sentiment analysis provide solutions. This report aims to provide an example of such a solution that can be used to derive knowledge from information and data presented in an unstructured format. The solution will aim to provide management information that will add value to the decision-making process of the organisation.

2.6 Cloud Computing

Cloud computing is the delivery of on-demand computing resources over the internet on a pay-for-use basis [86] from a provider. Some of the major providers include Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP) among others. Better known simply as the cloud, it offers benefits compared to the traditional enterprise that includes cost savings, flexibility, security, choice, speed, audit capabilities, integration, and business continuity opportunities. Recently the use of the cloud has increased and more and more services from public [87] and private, big [88] or small [89] are being offered through this delivery. The kinds of services offered through cloud computing are infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS). *Figure 3*, is a representation of the layers that make up the services and a definition of where the demarcation in terms of responsibility management

is determined (shown here as an example of AWS product). It is known as the shared responsibility model.

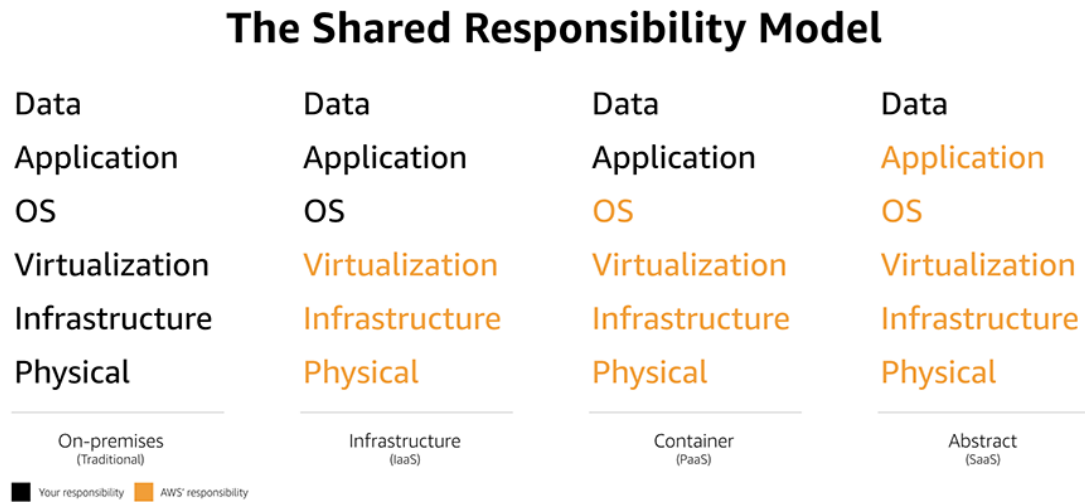


Figure 3 The Shared Responsibility Model [90]

2.6.1 Platform as a Service (PaaS)

Platform as a service is the service where a provider provides the hardware and software tools to its customers. *Figure 3*, shows that the customer responsibility is defined to be within the application and data layers while the provider handles the OS, virtualisation, infrastructure and the physical assets.

2.6.1.1 Function as a service (FaaS)

FaaS is a type of backend architecture where customers are provided with a complete platform just to run the code. The service provider manages the backend specifics of infrastructure. On such a service is AWS Lambda. AWS Lambda is an abstraction layer offered by AWS to allow for the running of compute functions. Rather than running virtual machine images or operating systems, the abstraction layer allows for the function

code to run while AWS provides the environment. The customer has no need to provide any infrastructure, no need to check resource status, security patches or updates. AWS Lambda also allows for event-driven execution of the code [91].

The advantages of FaaS are what defines the new architecture of cloud computing known as Serverless Architecture. Serverless architecture is a software design pattern that eliminates the need for managing infrastructure by the customer [92].

2.6.1.2 Database as a service (DBaaS)

DBaaS is a cloud service that offers the service of a database as part of PaaS. AWS RDS is an example of a DBaaS service that is offered by the provider AWS. It offers the services of popular database engines such as PostgreSQL, MySQL and Oracle [93].

2.6.1.3 Machine Learning as a Service (MLaaS)

Machine learning as a service (MLaaS) is a microservice that provides machine learning tools through cloud computing [94]. The integration of the models is done by using the prediction results through REST APIs. These services are currently being offered by the major cloud computing providers AWS Amazon ML, Microsoft Azure ML, IBM Watson, Google Cloud ML among others. It is also being offered as an add-on service by major consultancy companies such as Accenture [95], and Deloitte [96].

2.6.1.4 Business Intelligence as a Service (BIaaS)

Business intelligence as a service is a managed service that allows users to transform data that is held in different places into an actionable form of intelligence. This is done mainly through the use of visualisation. Amazon Quicksight [97] and Microsoft Power BI are examples of BIaaS. They allow the creation of dashboards for visualisation from different data sources e.g. files, databases, or data stores.

2.7 Datasets

A dataset is a collection of data of particular interest. In machine learning, datasets are used as the source of input in the development of ML models. The common datasets are those popularised through their use in published literature, journals or studies.

As interest in the AI and ML fields has increased the availability of datasets has increased too. It is now possible to find datasets online at various repositories such as Kaggle [98], KDNuggets [99], Github [100], public [101] and private [102], academic or commercial research labs [103].

2.8 Data Science

Data science is a discipline that provides principles, methodology, and guidelines for the analysis of data for tools, values or insights [104]. However, data science as a term has seen efforts to define it on what the scientists do. Conway [105] in an attempt to define the skills required in data science drew the data science Venn diagram.

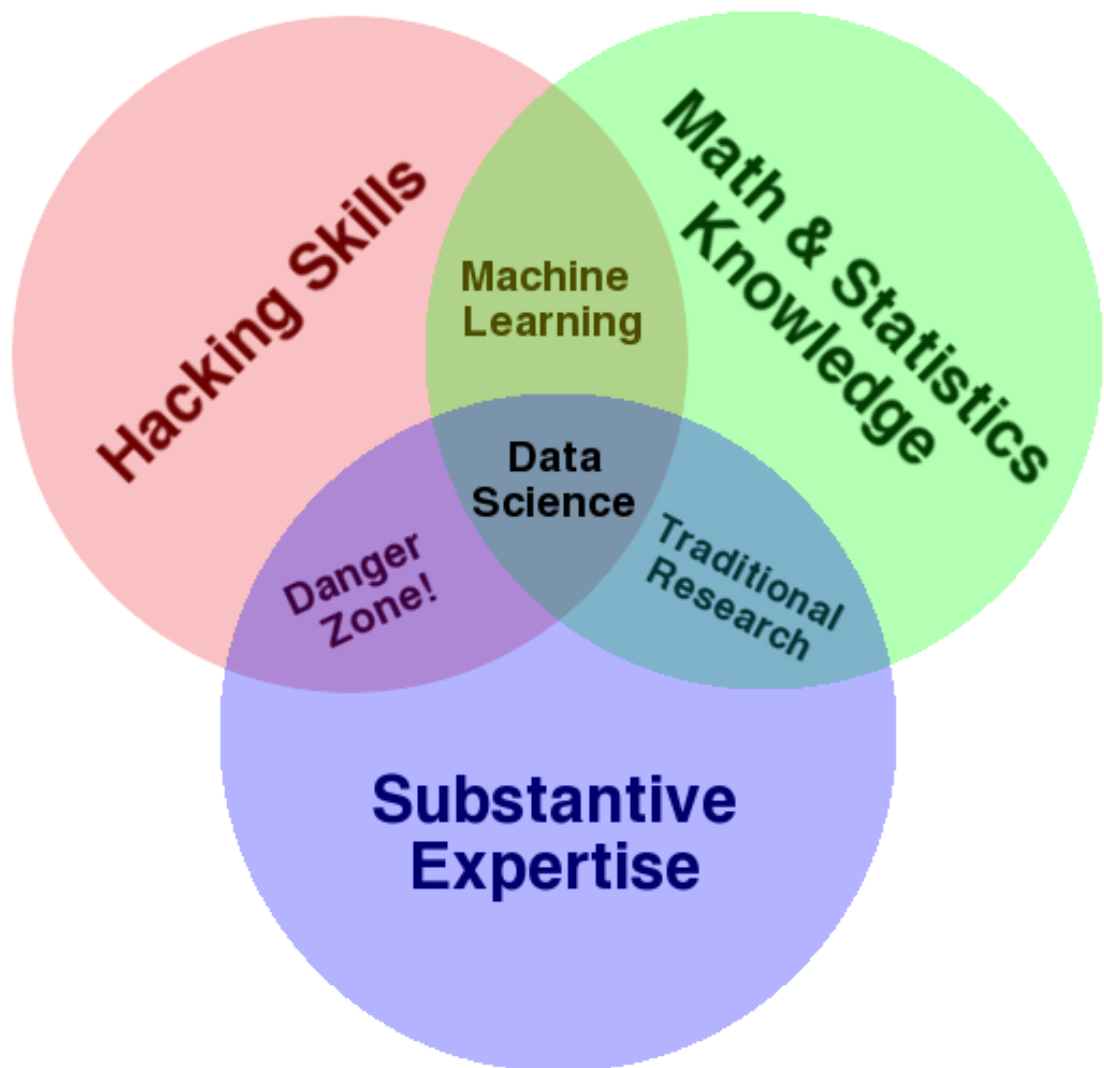


Figure 4 The Data Science Venn Diagram [105]

2.8.1 Software

2.8.1.1 Python

Python is an open-source usable high-level general programming language [106]. Currently, two versions of python are on general release and support, version 2.7 and version 3.8.1. Python is heavily used in machine learning due to its ease of use and easier learning curve relative to other programming languages. Many of the ML libraries and tools have been developed for use in Python and some have been integrated as inbuilt

libraries into Python. These include scikit-learn (used for data mining and analysis), Tensorflow (a high-level neural network library) and pylearn2 (used for data mining and ML).

2.8.1.2 Anaconda

Anaconda [107] is an open-source data science and machine learning platform that allows the use of Python and R software among other programming languages. The platform allows the management of libraries, dependencies and environments in one place through Conda. ML and deep learning models can be developed using tools such as scikit-learn, Tensorflow and Theanos. Jupyter notebooks are also part of the platform and provided as standard.

2.8.1.3 Jupyter Notebook

Jupyter notebook is an open-source web application that allows creation and sharing of documents that contain code, text, equations and visualisation [108]. The uses of Jupyter notebook include visualisation, data cleaning and manipulation, ML and other tasks that can be run through code.

2.8.1.4 NLTK

NLTK stands for Natural Language Toolkit. NLTK is a platform for building Python programs that involve working with human language data [109]. The toolkit was developed by the Department of Computer and Information Science at the University of Pennsylvania. The table, *Table 1*, gives a summary of the modules available for the specific natural language processing task. To make use of the modules, the modules have to be imported to the environment.

Table 1 NLTK modules for language processing tasks [109]

Language processing task	NLTK modules	Functionality
Accessing corpora	corpus	standardized interfaces to corpora and lexicons
String processing	tokenize, stem	tokenizers, sentence tokenizers, stemmers
Collocation discovery	collocations	t-test, chi-squared, point-wise mutual information
Part-of-speech tagging	tag	n-gram, backoff, Brill, HMM, TnT
Machine learning	classify, cluster, tbl	decision tree, maximum entropy, naive Bayes, EM, k-means
Chunking	chunk	regular expression, n-gram, named-entity
Parsing	parse, ccg	chart, feature-based, unification, probabilistic, dependency
Semantic interpretation	sem, inference	lambda calculus, first-order logic, model checking
Evaluation metrics	metrics	precision, recall, agreement coefficients
Probability and estimation	probability	frequency distributions, smoothed probability distributions
Applications	app, chat	graphical concordancer, parsers, WordNet browser, chatbots
Linguistic fieldwork	toolbox	manipulate data in SIL Toolbox format

2.8.1.5 Data Science Toolkit

Data Science Toolkit (DST) is an open-source collection of the best open data sets and open-source tools for data science, wrapped in an easy-to-use REST/JSON application programming interface (API) with command line, Python and Javascript interfaces [110]. The server can be run as a self-contained as a virtual machine.

2.9 Conclusion

In this section, we have discussed how artificial intelligence and machine learning has developed over time. We have also looked at what makes up natural language processing and how it is used in text analytics methods such as sentiment analysis and topic modelling through the use of LDA. We also identified places where datasets are available for use in machine learning projects. Finally, we looked at data science tools that are available and those that have will be used in the project.

3 Research Methodology

3.1 Introduction

Social media has seen exponentially grown as a form of organisational communication as highlighted in the Background Section in Chapter 1 and now represents an active channel for customer interaction. That interaction, however, is under the control of the customers and not the organisation and as such the form of that communication is “as is”. This project will use only publicly available data which will simplify and allow for working within the legal and ethical framework. Since the organisational institution is a public institution, the legal expectation is that any person within the UK should be able to access the services provided. The customer services for those services has an output that falls under the guise of unstructured data and would include the call centre voice calls, emails, chat logs, and social media interactions.

The aim of this project to provide a prototype text analytics cloud service that can offer sentiment analysis evaluation of the text input, provide topic guidance and provide a management information and business information output that can be used in the decision-making process of the customer service.

In this chapter, we will discuss the research philosophy used in the project. We will explain the research methodologies used and mention the outcomes of the research.

3.1.1 Philosophy

We decided to make use of positivism philosophy as we made use of real data captured from social media. We were independent of the parts of the organisation dealing with customer interaction and therefore holds an objectivist view. We look to provide a simplified view of the interaction from the customers as a dashboard. The sample used

involved the collection of over 200,000 individual social media posts captured over 10 years. The resultant prototype developed was developed using the population rather than making use of only a part of the data.

To better understand the current operation an introduction meeting was arranged to find out the true daily experiences of the teams of the analyst, support analysts, and management of the contact centre teams. This provided first-hand experience of the challenges experienced by the teams in terms of finding out what the customer was saying and how to identify any of the topical trends.

3.1.2 Approaches

The Literature Review chapter gives a systematic view of artificial intelligence, machine learning, text analytics, natural language processing, data, cloud computing and data science. We choose this view as it complements the use of the inductive approach.

The software development approach decided was to use prototyping for development. Use of high-fidelity prototypes (computer-based) in the approach allowed for the exploration of different iterations quickly and the testing of the functionality of the systems.

The deployment involved the inductive approach where the first development involved the use of Ubuntu chat logs to find out how Python, NLTK, data processing and topic modelling worked. Iteration of the model through the use of other natural processing tools in Python was used to further identify the preferred tool. The final prototype was developed in Azure ML using all the lesson learned during the iterative development.

3.1.3 Strategy

We made the use of grounded theory in the project. This offered the best approach for the project. It allowed for systematic and structured project development. We were able to

collect actual data that gave the true picture of what the consumers (public) were saying. Institutional knowledge allowed for the project to be able to identify the context that might have been lost on those without specific institutional knowledge.

3.1.3.1 Literature Review

The literature review was done to provide knowledge in the field of sentiment analysis and topic modelling. This informed the decision to pursue mixed methods approaches of how to develop the project including identifying the challenges offered by the data collected.

3.1.3.2 Design

We decided to use a sequential exploratory strategy where the collection of the data was done before any design of the model. The data influenced the choice of software to use and how to develop the model and finally on where best to deploy it. The prototype solution was required to meet the organisational strategic and operational plan for a cloud-first solution that includes the use of RESTful APIs.

3.1.4 Risks & Limitations

Due to making use of actual data, consideration of any legal implications had to be officially resolved. The requirements and time constraints were a factor as permissions were being clarified. To allow for the development process to start the use of a public dataset was determined to be a good substitute.

3.1.5 Validity, Reliability & Generalisability

3.1.5.1 Validity

The use of actual public data was necessary for the success of the project. The use of social media post was determined to be the validation of the prototype in deriving sentiment and topic modelling or classification.

3.1.5.2 Reliability

The use of public Twitter data allows any researcher to be able to replicate the results of this project.

3.1.5.3 Generalisability

The results from this project show that the use of social media text can be used to inform on sentiment. The use of cloud computing to deliver the prototype system implements it in the current working environment.

3.1.6 Data Evaluation Methods

3.1.6.1 Data Collection

To make use of chat logs, it was decided to make use of the public dataset available from the Ubuntu corpus as mentioned in the Introduction in the Focus section. The Twitter data was to be collected by running of a Python script.

During the evaluation of the system, 10 users were asked to rate 30 text phrases by sentiment scores of between -1 to 1 rating of negative to positive and decided on the number of topics they thought was contained in the text, see Appendix 12.4.1.

The sentiment analysis score and topic identification score were also collected in a PostgreSQL database for use in management information and business intelligence reporting.

3.1.6.2 Data Evaluation and Analysis

Once the system was deployed daily capturing of social media post from the DVLA social media accounts was undertaken with the results saved to a PostgreSQL database (a AWS RDS PostgreSQL database). The analysis of the resultant scores was presented through the use of a cloud service analytics tool, AWS QuickSight dashboard.

3.2 Development Methodology

3.2.1 Software Development Methodology

The software development methodology used was prototyping. This allows for instant feedback to be visible as the development progressed. The steps involved included identifying the requirements of the proposed system, the development of a prototype, a review of the system and iteration of the system to enhance it to capture the comments offered from the review. This methodology also suited the development team of one.

3.3 Resources

3.3.1 Hardware Resources

The hardware used for the project was a MacBook Pro i7.

3.3.2 Software Resources

The software used in the development of the project include:

- a) Python

As mentioned in the Software section in Chapter 2, Python is a high-level programming language in use commercially. Python offered the quickest learn-

to-development learning curve. The choice to use Python with machine learning project was obvious as the resources available made it the best choice.

b) NLTK

As mentioned in the Software section in Chapter 2, NLTK is a toolkit made for natural language processing tasks. NLTK was used for natural language processing tasks

c) MS Word

MS Word was used as the preferred word processing software for the project.

d) Mendeley

Mendeley was used to manage the reference collection.

e) MS Project

MS Project was used for project management planning.

f) Tweepy

Tweepy was used to get the tweets from Twitter API.

g) TweetScraper

TweetScraper was used to obtain tweets from Twitter Search without the use of the Twitter API.

h) Vader

Vader was used as a module in NLTK to undertake sentiment analysis.

i) Azure ML

We made use of Azure ML MLaaS for the final iteration of the prototype. Azure from the literature research was the most mature MLaaS for NLP tasks.

j) AWS Services

We made use of AWS services to fulfil the requirement of a serverless deployment for the prototype. The services used were AWS Lambda, API Gateway, AWS RDS (PostgreSQL) and Quicksight.

k) NodeJS

NodeJS was used to code the function code deployed in AWS Lambda for the connection to the Azure ML service and to persist the data to AWS RDS.

l) Bootstrap and JQuery:

Bootstrap and JQuery were used in the development of a web-based application that provided the structure for the consumption of the REST APIs developed. The version used was Bootstrap version 4 sourced from Bootstrap official source; <https://getbootstrap.com/>.

m) Postman:

Postman was used to test the Azure ML service endpoint.

3.3.3 Human Resources

The project involved a team of 10 users to rate the same piece of text used in the prototype for sentiment and topic valuation.

3.4 Conclusion

The research methodology choices were made to ensure that the project chooses the right method in data collection, and design of the prototype system. The choices made for development methodology informed what the software choice made to develop the design proposed. It also informed on the next step of requirement specification of the proposed prototype.

4 Software Requirements Specification (SRS)

4.1 Introduction

This section is to define the software requirements of the prototype system developed as part of this report.

4.2 Goals of SRS:

The final system developed requires an input of unstructured data in the form of texts from social media posts and then to return a sentiment score and predictive topics. The system should be hosted on the cloud and use resources as practically possible. The system should provide RESTful APIs, where the resources should be accessed through a defined endpoint. A dashboard should be provided as part of the final system to provide a visual representation of the system results.

4.3 Fundamentals

4.3.1 Functional Requirements:

The functional requirements are those descriptions that the system should do [111].

Table 2 Example of Functional requirement

Name	F1: Allow input of text
Summary	The system should allow the inputting of copied text.
Rationale	An analyst will copy the text of a social media post and paste it into the system to carry out sentiment analysis and topic identification.
Requirement	When the user starts the system the user should be able to see where to enter copied text.
Reference	UC-1: Enter text

4.3.2 Non-functional Requirements:

Non-functional requirements are software qualities attributes of the system [111].

Table 3 Example of Non-functional requirement

Name	NF-1 Availability of the system through a web-browser
Summary	The system should be available through a web-browser for convenience.
Rationale	Since all analyst uses the web-browser for other activities it is best to offer the service through the web-browser.
Requirements	Ensure the service works in all available web browsers.
References	UC-1: Enter text

The attributes that form part of non-functional requirements include availability, efficiency, flexibility, portability, integrity, performance, reliability, reusability, robustness, scalability, usability [111].

4.3.3 Use-cases:

Use cases are a representation of high-level functionality of a system such as a component part.

Table 4 Example of a use case

Name	UC-1 Enter Test
Summary	Text data can be entered
Rationale	
Users	Analyst
Preconditions	The user has already copied the text from the source.
Basic course of events	1. The user activates the text area 2. The user can enter the copied text into the area
Alternate path	None
Postcondition	If the user does not want to proceed with the text inserted they can clear the text area

4.4 Process

The actors of the prototype system are:

Table 5 Actors of the prototype

Name	Responsibility
Analyst	Enter the text into the system and run the sentiment analysis and topic identification
Report Analyst	Be able to get management information and business information from a dashboard
Team Manager	Be able to easily know from prepared reports the sentiment expressed by customers.

Figure 5 shows the sequence diagram of how an analyst uses the system. It shows the process of getting the sentiment analysis score and topic identification probability percentages and saving the information to a database.

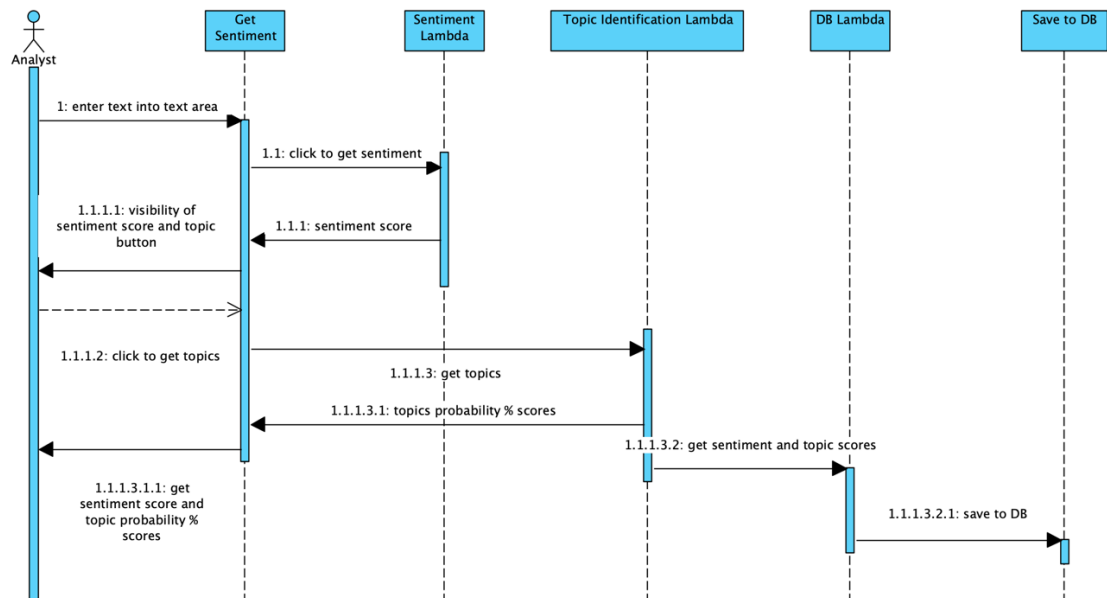


Figure 5 Sequence diagram of the prototype with the analyst actor

4.5 Validation

The validation of the prototype system will be done using social media posts captured from Twitter and Facebook posts and results evaluated in the Results and Evaluation section.

5 Design and Development

5.1 Design

The most important thing is to ensure that the developed system gives value to the organisation. Michalewicz et.al [112] noted that successful management decisions are made when the decision makers are able to leverage the ability of knowledge offered if accompanied with some recommendations. One of the aims of the project is that once the predictions of the topics and sentiment derived from the entered text the results are saved where a management information dashboard can be used to offer some real-time information in a useful manner.

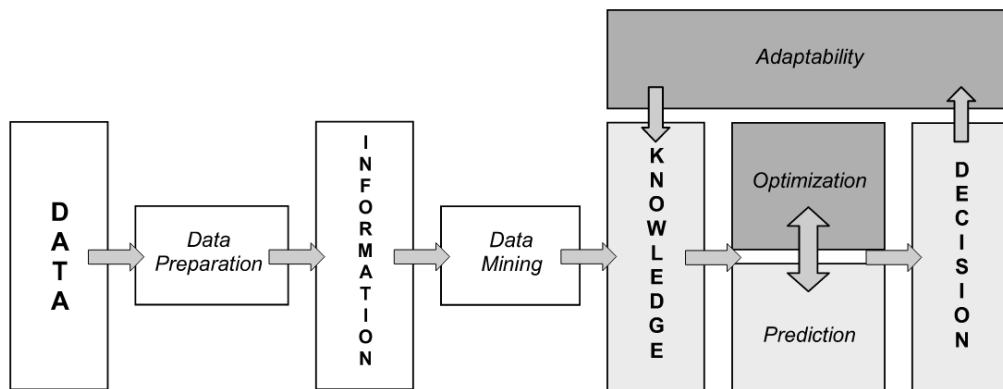


Figure 6 Use of prediction and optimisation to add value in business systems [112]

The *Figure 6* above shows an overview of the transformation that data and derived information need to undergo through text mining process to achieve some knowledge that provides value in decision making process through the prediction and optimisation provided by the developed system. The developed system will be a functioning text analytics system that will provide sentiment and topic prediction.

We used the unstructured data found in the Ubuntu Dialogue Corpus [16] dataset to develop the methodology in terms of topic modelling. Once this was done, DVLA Twitter

messages was then used to refine the topic modelling and add sentimental analysis to the system.

5.1.1 Design Outlook

The overall design flowchart of the system is as shown in *Figure 7* below.

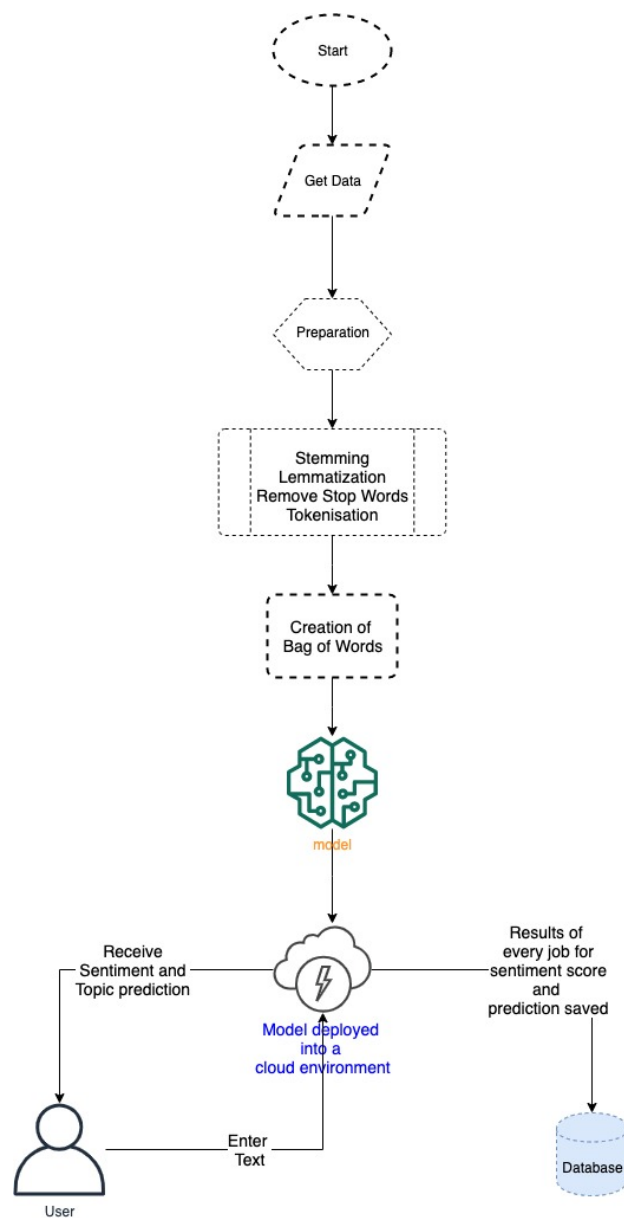


Figure 7 Prototype system design

5.1.2 Data Collection

5.1.2.1 Ubuntu dataset

The Ubuntu corpus dataset was downloaded from the research repository [113]. The data includes ubuntu chat support logs of a 1-to-1 conversation.

A	B	C	D	E	F
folder	dialogueID	date	from	to	text
301	1.tsv	2004-11-23T11:49:00.000Z	stuNNed		any ideas why java plugin takes so long to load?
301	1.tsv	2004-11-23T11:49:00.000Z	crimsun	stuNNed	java 1.4?
301	1.tsv	2004-11-23T11:49:00.000Z	stuNNed	crimsun	yes
301	1.tsv	2004-11-23T11:49:00.000Z	crimsun	stuNNed	java 1.5 loads _much_ faster
301	1.tsv	2004-11-23T11:50:00.000Z	stuNNed	crimsun	noneus: how can i get 1.5 is there a .deb somewhere?
301	1.tsv	2004-11-23T11:50:00.000Z	crimsun	stuNNed	not yet.
301	1.tsv	2004-11-23T11:50:00.000Z	stuNNed	crimsun	noneus: is this blackdown or sun?
301	1.tsv	2004-11-23T11:50:00.000Z	crimsun	stuNNed	did you install just the jre?
301	1.tsv	2004-11-23T11:51:00.000Z	crimsun	stuNNed	i use IBM's 1.4.2
301	1.tsv	2004-11-23T11:51:00.000Z	crimsun	stuNNed	(jdk, because i do globus development)
301	1.tsv	2004-11-23T11:52:00.000Z	stuNNed	crimsun	ah ok then, i'll give it a try, does IBM have a standard jre with plugin?
301	1.tsv	2004-11-23T11:52:00.000Z	crimsun	stuNNed	yes

Figure 8 Original Ubuntu dataset

Due to the large size of the dataset, we only made use of conversations contained in the “ubuntu_dataset.tgz.aa (954M)” file.

5.1.2.2 DVLA dataset

Using TweetScraper which is mentioned in Software Resources section we obtained tweets associated with the @DVLAgovuk, the official DVLA Twitter handle [114], and #DVLA tweets. TweetScraper is a simple crawler for Twitter Search without using the Twitter API [115]. Each individual tweet is saved as a JSON file.

```
{
  "usernameTweet": "TheABD",
  "ID": "1330565166",
  "text": "ABD Welcomes MPs` Defence Of Drivers` Privacy - http://www.abd.org.uk/pr/656.htm #Parking #DVLA",
  "url": "/TheABD/status/1330565166",
  "nbr_retweet": 0,
  "nbr_favorite": 0,
  "nbr_reply": 0,
  "datetime": "2009-03-15 06:56:16",
  "is_reply": false,
  "is_retweet": false,
  "user_id": "21440544"
}
```

Figure 9 Example of TweetScraper tweet file

5.1.3 Data Pre-processing

Before data can be used in an AI process, data requires to be processed. During processing the data is cleaned to remove unnecessary content that rather than inform will be a hinderance. In the best-case scenario, the hinderance can provide a false positive outcome, which is where the outcome as an output of the system is deemed to positively answer the questions posed. An example of this would be if an organisation is looking to find out if the new payment system deployed is being well received. If the answer from the process is no while it is yes, then the organisation will spend more time improving the system to meet the dissatisfaction. In statistical terms this is known as a type I error. And in the worst-case scenario, it could be the output saying the process has been well received while in reality the organisation is disappointing customers and destroying its reputation and losing customers and money in the process. This in statistical terms would be known as a type II error.

5.1.3.1 Ubuntu dataset

For pre-processing, we had to change the structure of the dataset to ensure that data was one line per conversation. We used Python with the help of pandas to read the dataset file and merge the conversations by grouping the ID, see Appendix 12.2.1.1.

5.1.3.2 DVLA dataset

To do this, we used a Python with the help of pandas to read all the individual files from the saved location and merge them into a single csv file, see Appendix 12.2.1.2.

5.1.3.3 Load data

To load the data into the Python environment, we used pandas dataframes as shown in *Figure 10*.

```
import pandas as pd
ubuntu_chat = pd.read_csv('Merge.csv')
```

Figure 10 Example of importing data

5.1.3.4 Normalisation

Normalisation is the conversion of text into only lowercase. This removes any distinction of words due to the cases on the letters, shown in *Figure 11*.

```
# Normalise the text
ubuntu_chat['text'] = ubuntu_chat['text'].str.lower()
```

Figure 11 Example of normalisation

5.1.3.5 Tokenisation

Tokenisation is the process whereby text that is made up of words, sentences, or clauses is split into separate single parts. It involves the removal of punctuation marks in the texts and in the most basic form gets the text into just words.

```
# Tokenise the text
ubuntu_chat['words'] = ubuntu_chat['text'].apply(nltk.word_tokenize)
```

Figure 12 Example of tokenization

5.1.4 Stop Words

Stop words are removed from data since they are unlikely to be of any benefit and will get in the way of undertaking natural language processing (NLP) task of text classification. It is however advised not to undertake the removal of stop words in NLP tasks such as machine translation and text summarisation, since the use of the stop words in this case does inform the feel and gives meaning to the text. We make use of NLTK modules to import English stopwords. It is not ideal to only use English as we know the interaction with the DVLA can be in English or Welsh.

```
# Load stopwords choice of using nltk stopwords
from nltk.corpus import stopwords
stop_words = stopwords.words('english')
# Remove stop words from the text
ubuntu_chat['words_minus'] = ubuntu_chat['words'].apply(lambda x: [item for
item in x if item not in stop_words])
```

Figure 13 Example of stop words removal

5.1.4.1 Stemming

Stemming is the process where a word is mapped to its base form in the language, a stem. It involved the use of mainly either removing the suffixes or prefixes. Stemming is therefore a necessary part of any natural language processing activity. Singh and Gupta [116] provide a categorisation of stemming techniques in *Figure 14* and highlight the different ways and algorithms to find the stems of words in natural language processing (NLP). However, it is possible to get errors through either over-stemming or under-stemming. Over-stemming would happen when two words are associated to the same base word but should not be and under-stemming happens when for example two words are associated with different stems but they are part of only one stem.

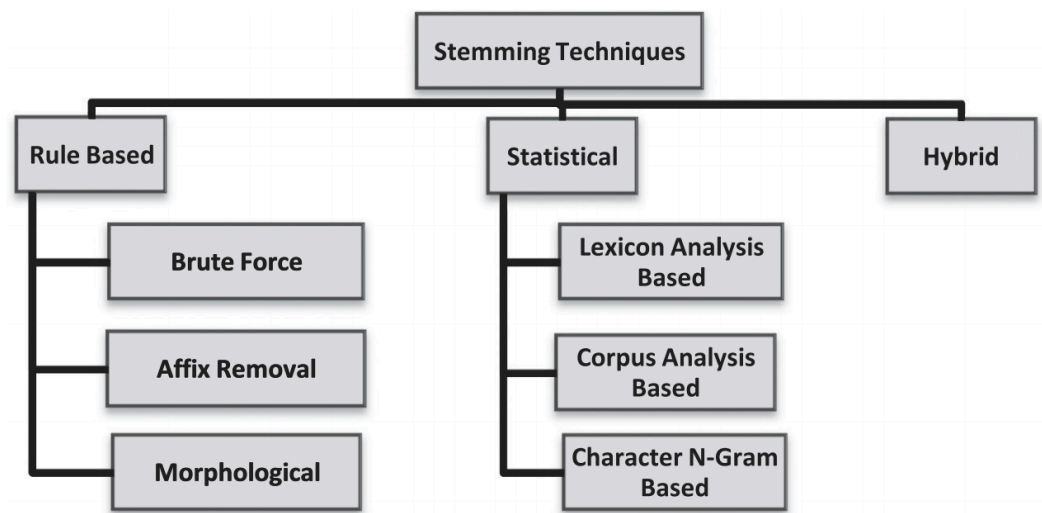


Figure 14 Stemming Techniques [116]

An example of the stemming done to the data is shown in the figure below.

```
#Snowball Stemmer
from nltk.stem.snowball import SnowballStemmer
stemmer = SnowballStemmer("english")
ubuntu_chat['words_stemmed'] = ubuntu_chat['words_minus'].apply(lambda x:
[stemmer.stem(y) for y in x])
```

Figure 15 Example of stemming

5.1.5 Extracting features

The simplest method of extracting features from text is making use of modules that are part of the toolkit. For example extracting parts of speech from the text would require the use of “pos_tag” module in the NLTK, as shown in *Figure 16*.

```
# extract part of speech from the stemmed column
ubuntu_chat['pos']= ubuntu_chat['words_stemmed'].apply(nltk.pos_tag)
```

Figure 16 Example of extracting parts of speech features

5.1.6 Data Visualisation

Data visualisation involves use of tools like graphs and maps to represent the data in a visual way. Data visualisation tools are available as part of the Python language or add-on modules. They have become an integral part of machine learning project as visual aids work better than printed results. Example of data visualisation tools include matplotlib, a Python 2D plotting library [117], and pyLDAvis [118].

5.1.7 Modelling

Model building is a process of reification, where more details is being added to the overall model during implementation. The projects aim to deliver sentiment analysis and topic identification but how? We decided on making use of using Vader sentiment module in NLTK and using LDA for topic modelling.

Following the overview given in *Figure 17*, the modelling process through the iterations will be used to refine the results.

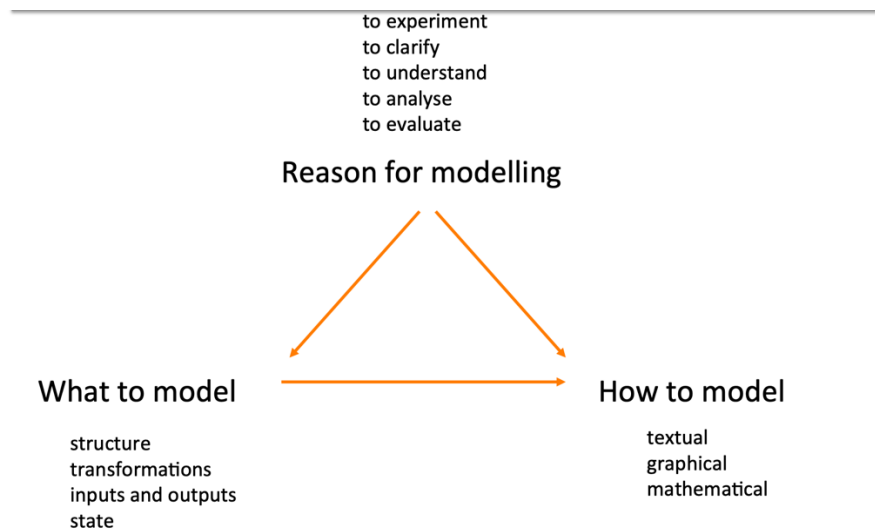


Figure 17 Modelling overview [119]

In NLTK the sentiment analysis was done using Vader Sentiment that is part of the sentiment class. Vader [120] is a simple rule-based model for general sentiment analysis that was shown to outperform individual human raters. An example of how Vader was used is shown in *Figure 18*.

```
# Sentiment Analysis using Vader
from vaderSentiment import vaderSentiment as vs
sentimentAnalysis = vs.SentimentIntensityAnalyzer()

# function for undertaking the analysis
def sentiment_analyzer_scores(sentence):
    score = sentimentAnalysis.polarity_scores(sentence)
    print("{:-<40}{:}".format(sentence, str(score)))

# getting the scores for a line of text in the dataframe
sentiment_analyzer_scores(ubuntu_chat['tolist'][0])
```

Figure 18 Example of using Vader in NLTK

For topic modelling, we decided after a review when undertaking Literature Review that we consider the use of non-negative matrix factorization (NMF) and LDA using Gensim. A review of using NMF is shown below. It was then decided to use Gensim for topic modelling.

```
#use sklearn in NLTK for NMF a
from sklearn import decomposition
nmf = decomposition.NMF(init = 'nndsvd', n_components = 2, max_iter = 200)
W = nmf.fit_transform(tfidf)
H = nmf.components_
feature_names = vectorizer.get_feature_names()
n_top_words = 10
# print top words in each topic
for topic_idx, topic in enumerate(H):
    print("Topic #%d:" % topic_idx)
    print(" ".join([feature_names[i]
                    for i in topic.argsort()[:n_top_words - 1:-1]]))
    print()
# results
Topic #0:
ubuntu use help anyon know work tri window hi file
Topic #1:
instal apt packag sudo ubuntu updat cd upgrad desktop remov
```

Figure 19 NMF using NLTK

5.2 Development

The development prototype iterations were done with *Figure 7* as the reference.

5.2.1 ML model

5.2.1.1 Iteration 1

Iteration 1 was developed in NLTK, using the Ubuntu chat logs dataset to experiment on how to import modules, import data, undertake data pre-processing, and understand how to display the most talked terms in the chats.

Results of the model for the most talked terms in the dataset is shown in *Figure 20*.

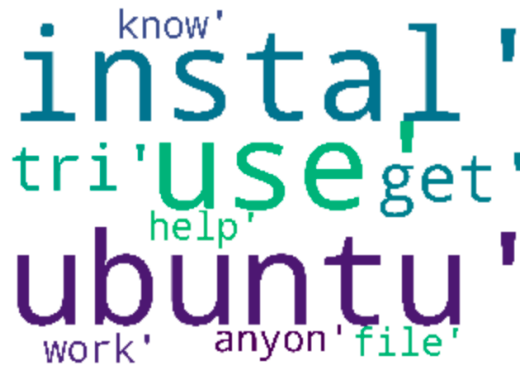


Figure 20 Iteration 1 results

5.2.1.2 Iteration 2

This iteration built on the foundations of Iteration 1 while making changes on the feature extraction adding the use of NMF to identify topics relevant in the Ubuntu dataset. The results shown in *Figure 21* are the result of running NMF analysis to find 2 topics from the dataset.

Topic #0:
ubuntu use help anyon know work tri window hi file

Topic #1:
instal apt packag sudo ubuntu updat cd upgrad desktop remov

Figure 21 Iteration 2 NMF results

5.2.1.3 Iteration 3

This iteration was done to add sentiment analysis to the model previously achieved in Iteration 2. The result of the iteration is shown in *Figure 22*.

```
sentiment_analyzer_scores(ubuntu_chat['tolist'][0])
[{'', 'also', 'guy', ',', '"m', 'tri', 'get', 'firefox', 'prefer', ',', 'keep', 'tell', 'root.dis', '=', 'true', '"',
';', 'log', 'root', '"', '?', '.', '}' ]{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
```

Figure 22 Iteration 3 Sentiment Analysis

5.2.1.4 Iteration 4

Having achieved the objectives of getting sentiment and topic identification it was agreed to proceed to using the DVLA dataset. And following the Literature Review on topic modelling, in this iteration we decided find out more about topic modelling using NMF and LDA.

```
# Topic modelling using NMF and LDA
from sklearn.decomposition import NMF, LatentDirichletAllocation
```

Figure 23 NMF and LDA modules

The NMF was done using the Frobenius norm, and using the generalized Kullback-Leibler divergence. The results for 8 topics for the two NMF models and the LDA model are shown in *Figure 24*.

```
Fitting the NMF model (Frobenius norm) with tf-idf features, n_samples=2000 and n_features=1000...
done in 15.360s.

Topics in NMF model (Frobenius norm):
Topic #0: contact details hi dv la team advice investigate gov uk
Topic #1: dvlagovuk license address just car change new help sent ve
Topic #2: thanks ok help reply great hi dm dvlagovuk sorted able
Topic #3: licence driving hi old return address dvla sa99 swansea cut
Topic #4: thank ok dvlagovuk great advise able reply help okay sorted
Topic #5: online enquiry 123 0815 0300 team hi contact ll advice
Topic #6: com twitter pic uk gov dvla driving scam tell government
Topic #7: tax vehicle car report untaxed hi uk gov road v5c

Fitting the NMF model (generalized Kullback-Leibler divergence) with tf-idf features, n_samples=2000 and n_features=1000...
done in 93.105s.

Topics in NMF model (generalized Kullback-Leibler divergence):
Topic #0: hi contact details gov advice uk need number team information
Topic #1: dvlagovuk car just dvla know got does ve new drive
Topic #2: dvlagovuk thanks help dm able just ok great sent reply
Topic #3: licence hi address driving change need application license apply new
Topic #4: dvla driving thank licence drivers counterpart swansea changes news drivinglicence
Topic #5: dvlagovuk tax online car 0300 service need 123 ca 0815
Topic #6: com pic twitter gov uk dvla check driving tell apply
Topic #7: hi vehicle dvla tax send report untaxed aware message delete

Fitting LDA models with tf features, n_samples=2000 and n_features=1000...
done in 388.862s.

Topics in LDA model:
Topic #0: com twitter pic advise helps dvlagovuk stat worry surely country
Topic #1: texts emails com scam pic taxitorloseit information matter dvla ly
Topic #2: licence hi driving address dvlagovuk vehicle send uk need new
Topic #3: contact hi details team advice la dv enquiry dm 0300
Topic #4: dvlagovuk thanks thank tax car vehicle twitter ve service just
Topic #5: delete social media twitter com parking autism isn directly unable
Topic #6: sa99 follow lzw dvla com asked assistance lol umber visitalk
Topic #7: record afraid tweet com things response twitter clear pic autistic
```

Figure 24 Results of running NMF and LDA

5.2.1.5 Iteration 5

Following further research, it was acknowledged that LDA can be improved. It was then decided to make use of Mallet [121], a JAVA based package for statistical natural language processing and topic modelling among other tasks. To achieve this objective, it was agreed to use Gensim, which is a library in Python [122] [123]. To use Mallet, we employed the use of the Gensim python wrapper available [124].

The coherence score is used to rate the quality of the topics derived by the LDA model [125]. *Figure 25*, shows using only LDA and *Figure 26* shows the model after using LDA Mallet.

Coherence Score: 0.2976228777140434

Figure 25 LDA as normal

Coherence Score: 0.4699788388530619

Figure 26 LDA after using LDA Mallet

The topics summaries are shown as word clouds in *Figure 27*.

Figure 28 show the topics as defined by the model after improvement through LDA Mallet and where the probability is equal across the topics. It also shows the most salient terms in the dataset. Further detail of the individual topics including the words that are most salient is available in the Appendix section 12.2.1.3.

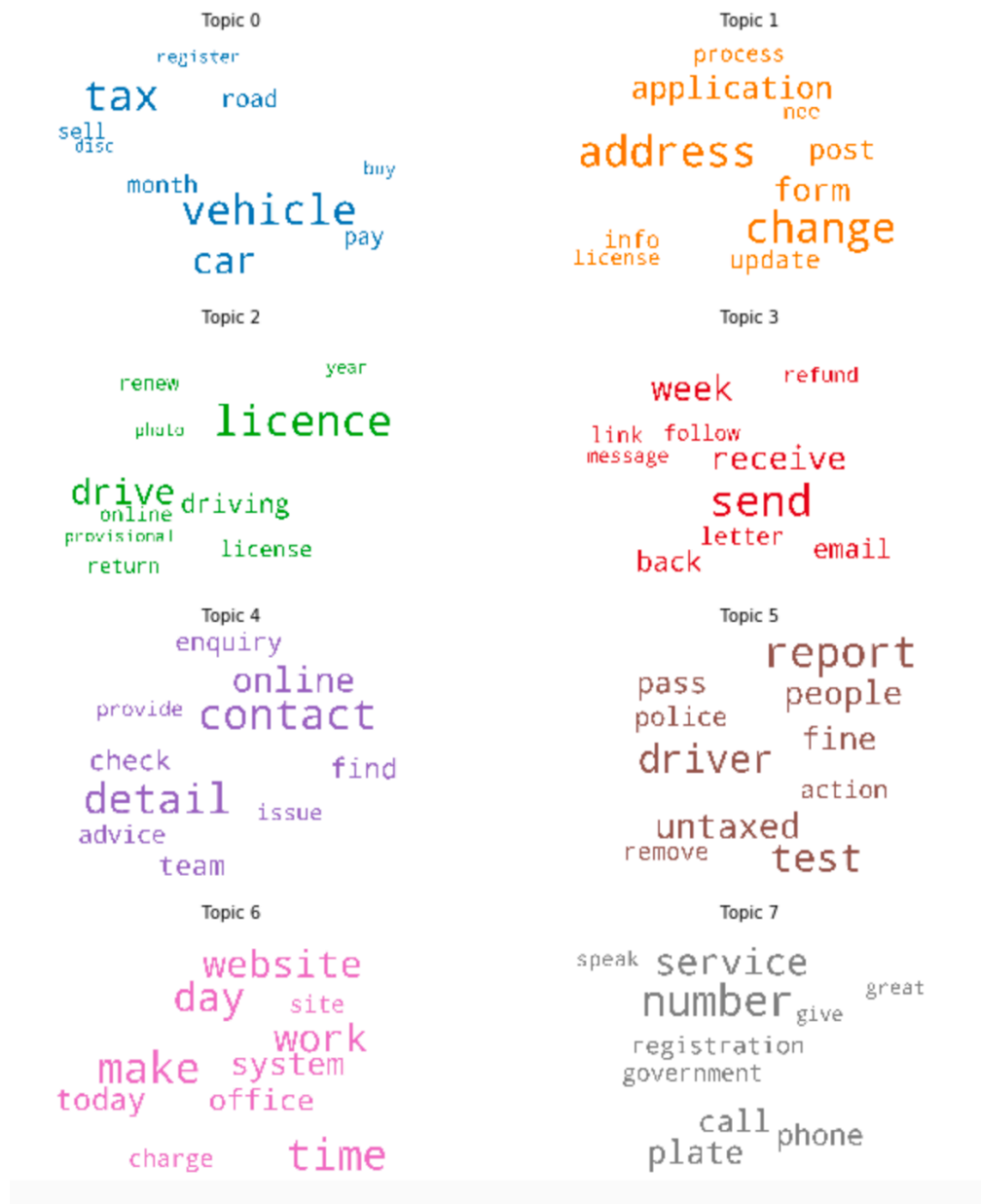


Figure 27 Topics from iteration 5

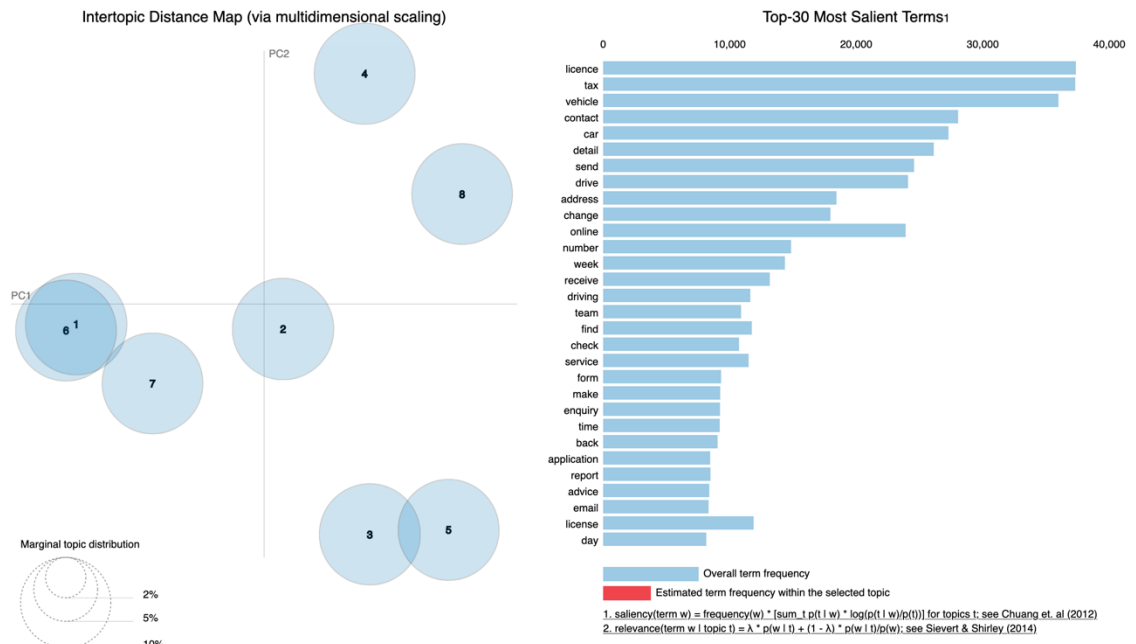


Figure 28 Topic modelling results

5.2.1.6 Iteration Azure ML

Having developed the model in Python, and NLTK, and established the concepts, it was time to start development in the cloud. Of the three major cloud providers i.e. Microsoft Azure, Amazon Web Services and Google Cloud Platform, Microsoft Azure provided the most developed machine learning natural language platform.

We choose to develop the model in Microsoft Azure Machine Learning studio (classic) (AMLS). To start, we uploaded the dataset through the graphic user interface (GUI), as shown in Figure 29.

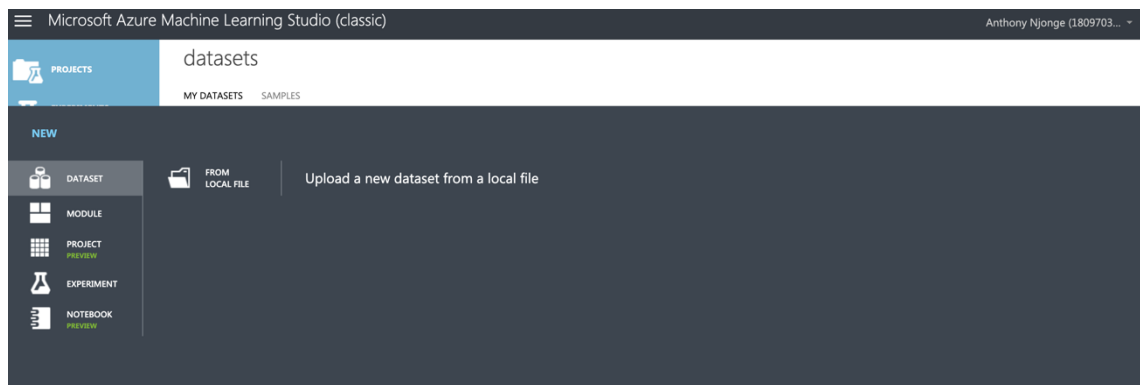


Figure 29 Azure dataset upload

Using the studio was intuitive as the tasks further required were the same as the Python and NLTK experience with a GUI provided, as shown in *Figure 30*.

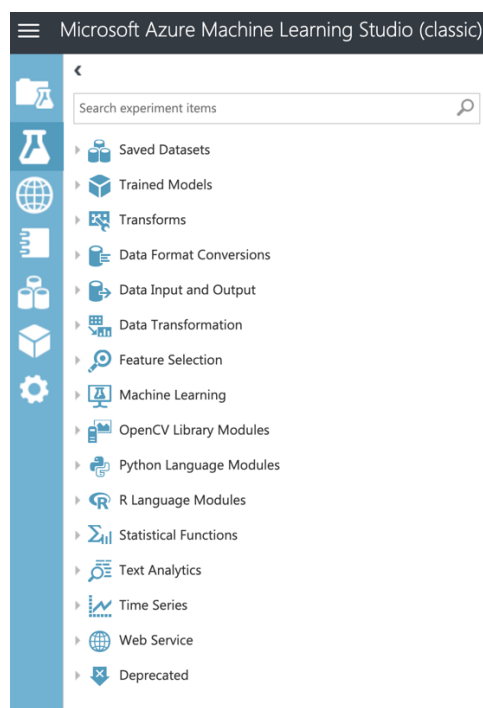


Figure 30 Machine learning development tasks

To start the pre-processing of the dataset, the studio provides “Preprocess Text” option that requires further fine tuning depending on requirement, as shown in *Figure 31*.

The screenshot shows the 'Properties' tab of the AMLS interface. Under the 'Text column to clean' section, a list of columns is shown with 'Text' selected. Below this, a 'Launch column selector' button is present. A list of 15 text processing options is displayed, all of which are checked:

- Remove stop words
- Lemmatization
- Detect sentences
- Normalize case to lowercase
- Remove numbers
- Remove special characters
- Remove duplicate characters
- Remove email addresses
- Remove URLs
- Expand verb contractions
- Normalize backslashes to slashes
- Split tokens on special characters

Figure 31 AMLS Pre-processing

We further choose LDA options and K-means clustering to create the required model in AMLS as shown in Figure 32.

The screenshot displays two configuration panels side-by-side. The left panel is for 'Latent Dirichlet Allocation' (LDA) and the right panel is for 'K-Means Clustering'.

Latent Dirichlet Allocation (LDA) Options:

- Target column(s): Selected columns: Column names: Preprocessed Text
- Number of topics to model: 8
- N-grams: 2
- Normalize: ☒
- Show all options: False
- Build dictionary of ngrams: True
- Maximum size of ngram dictionary: 20000

K-Means Clustering Options:

- Create trainer mode: Single Parameter
- Number of Centroids: 3
- Initialization: K-Means++
- Random number seed: 12
- Metric: Euclidean
- Iterations: 100
- Assign Label Mode: Ignore label column

Figure 32 LDA and K-means options in AMLS

The final training model including options for webservices input and output is shown in Figure 33. It was then deployed as a webservice as *Figure 34*. The service provides an API endpoint that was then available to be integrated with any other service.

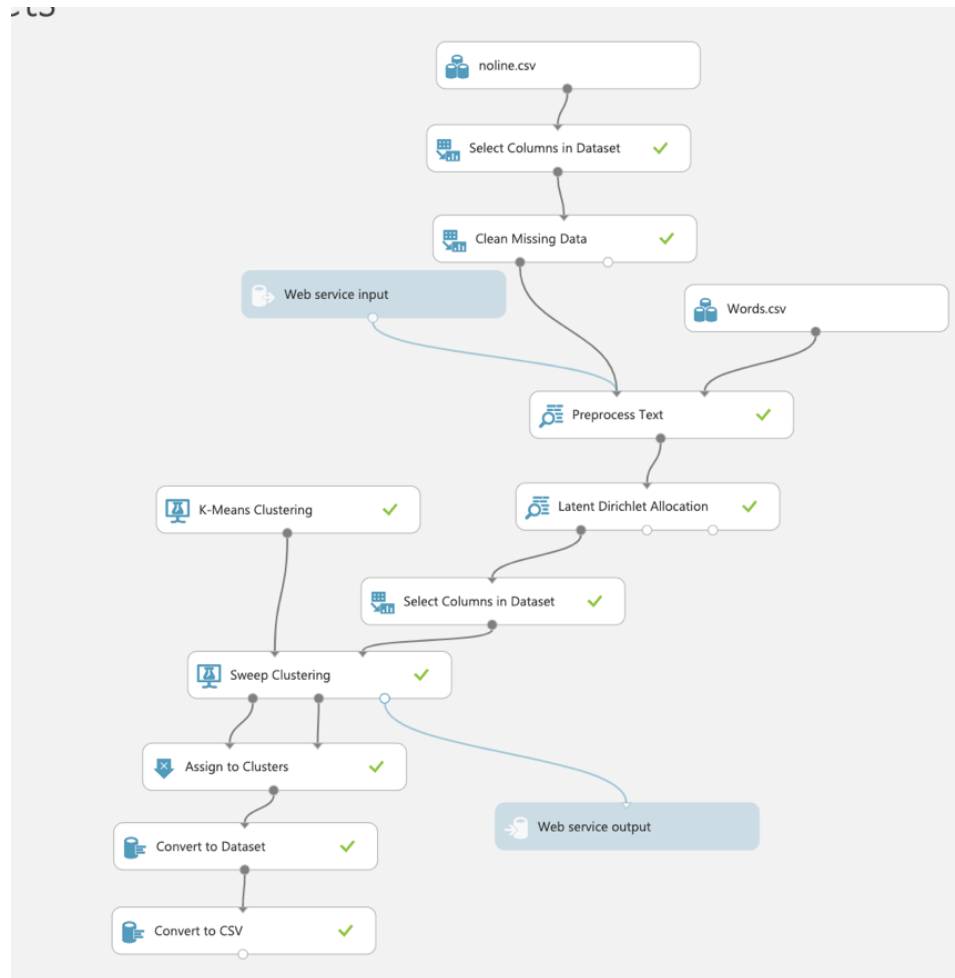


Figure 33 AMLS final training experiment

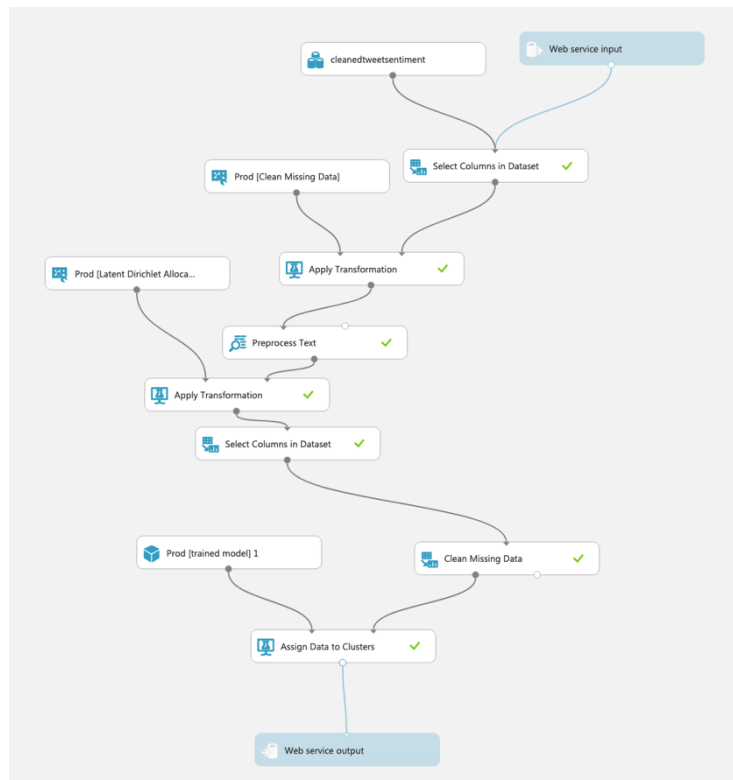


Figure 34 AMLS predictive service

Using Postman, we were able to test the endpoint, as shown in the Figure 35.

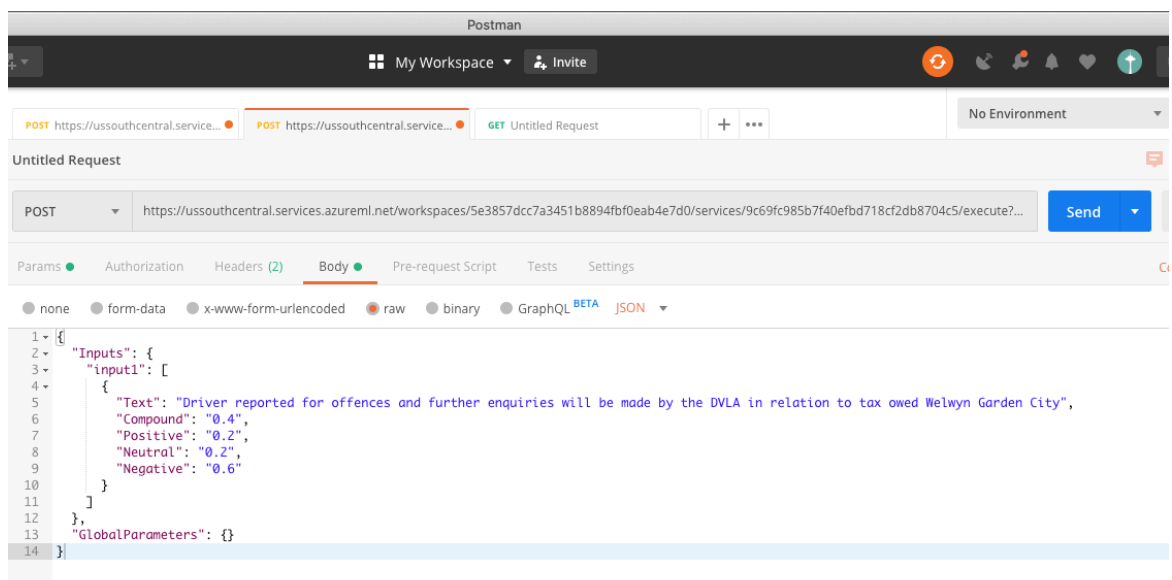


Figure 35 Testing webservice endpoint with Postman

5.2.2 Production API

To meet an objective of the project to ensure the service is provided through cloud serverless technologies. All services were to be delivered through APIs. The service was planned to be delivered as shown on *Figure 36*. We decided to use the AWS services for the delivery of the APIs. Amazon API Gateway enables the creation and deployment of developed REST and WebSocket APIs at any scale. We choose to use AWS Lambda for FaaS and AWS RDS for data storage of the results.

The service was to make use of AWS Lambda with two NodeJS functions for the Azure ML endpoint (*Figure 38*) and the PostgreSQL database interaction (*Figure 39*). And one AWS Lambda Python function for the Sentiment Analysis, *Figure 37*.

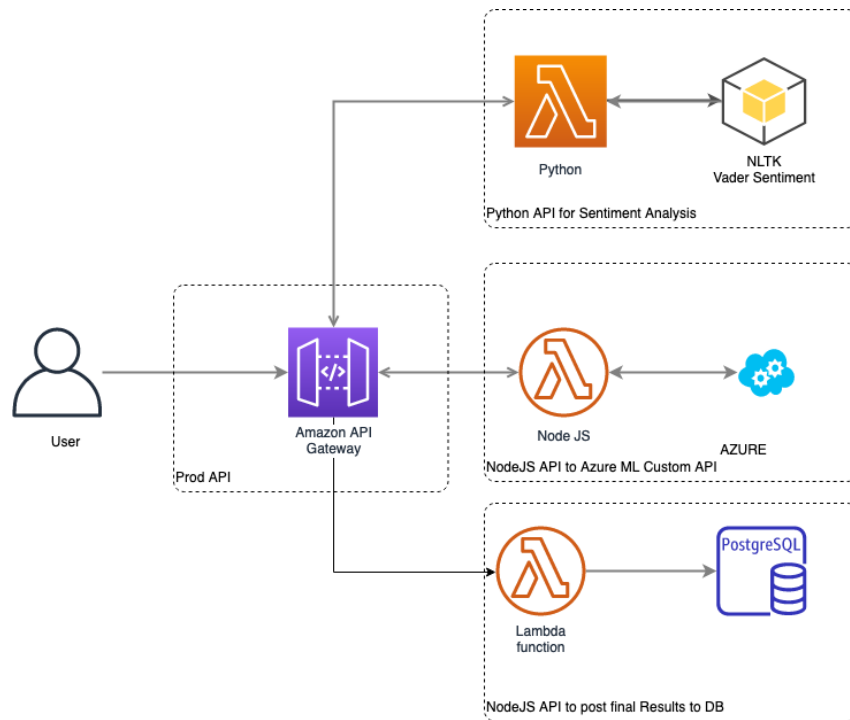


Figure 36 Deployed Service

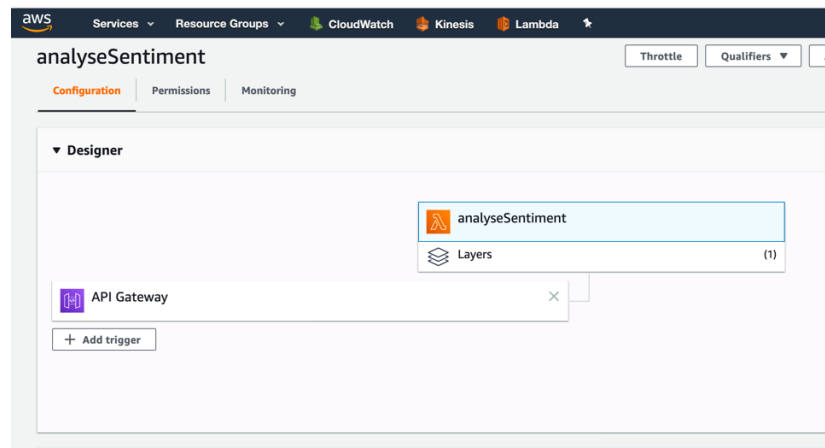


Figure 37 Sentiment lambda

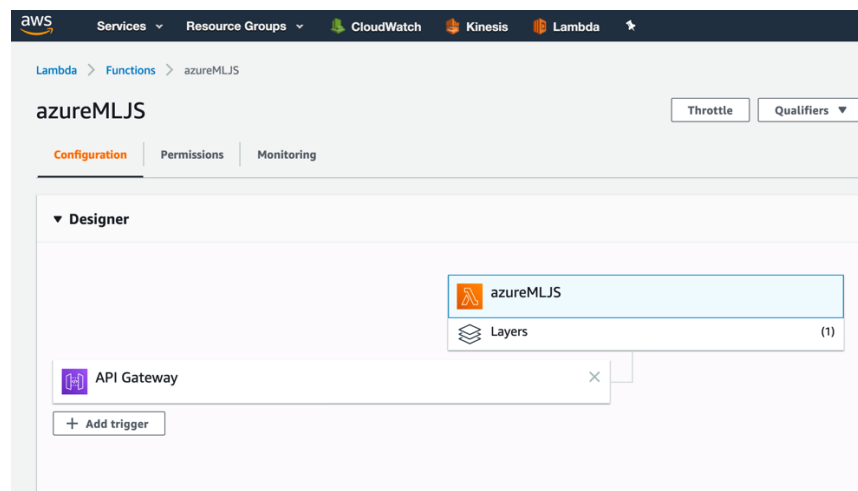


Figure 38 Azure lambda

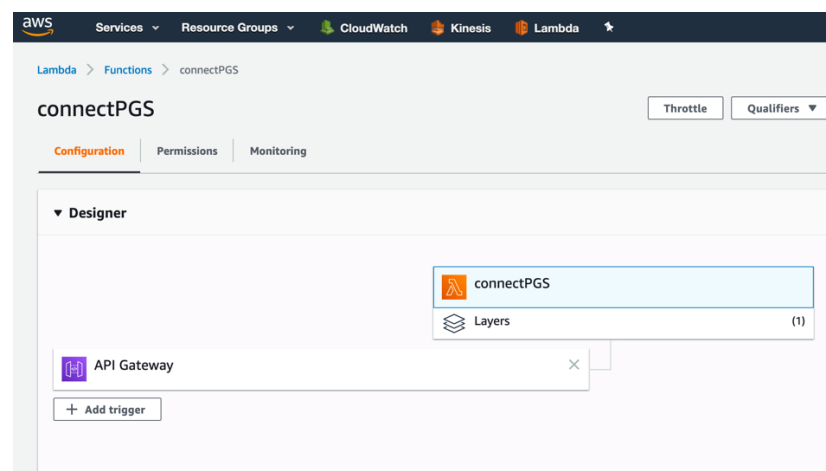


Figure 39 PostgreSQL lambda

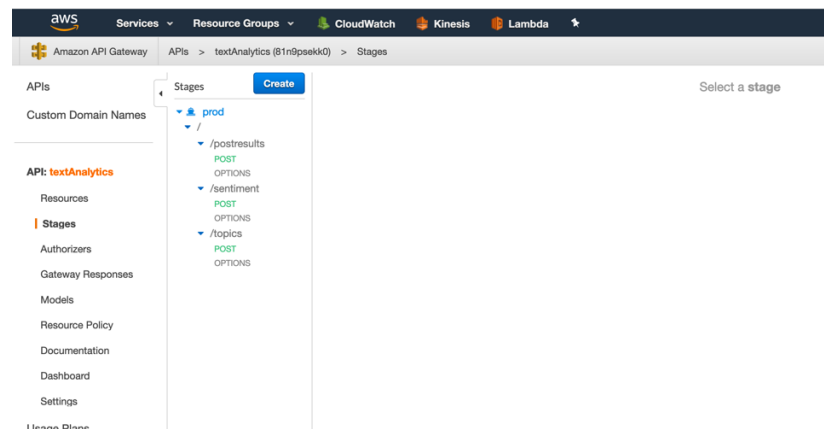


Figure 40 Amazon API Gateway

5.2.3 Database Development

For the database, we decided to make use of the Amazon RDS services with the choice of a PostgreSQL database.

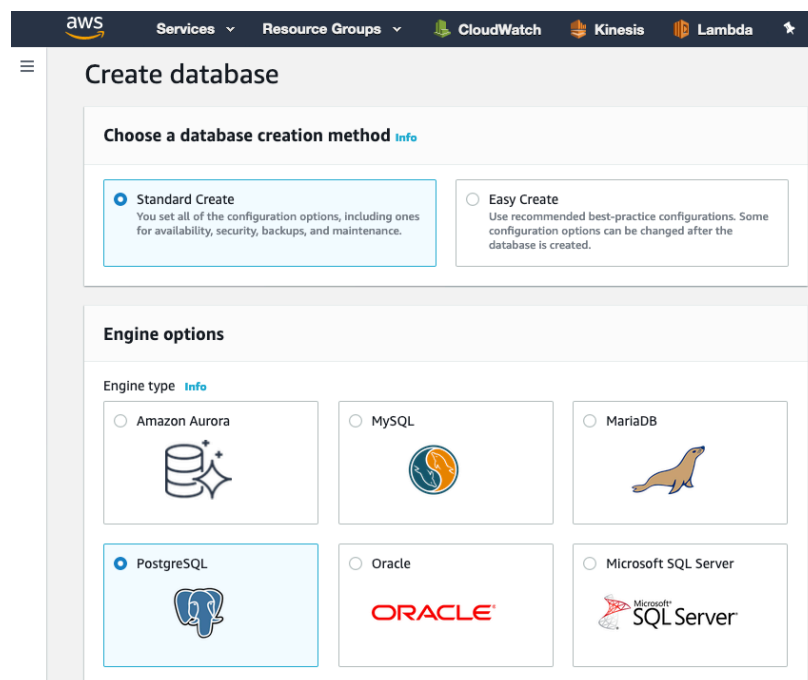


Figure 41 Amazon RDS PostgreSQL database

5.2.4 UI Development

The user interface (UI) was designed using Bootstrap which was highlighted in the Software Resources section. We choose Bootstrap as was developed for use in web application development prototyping. We used JQuery to provide functionality for the connection from the user interactions e.g. clicks. The Javascript functions available are to connect to the AWS API Gateway endpoint that provides the score results and posting of the data to the database.

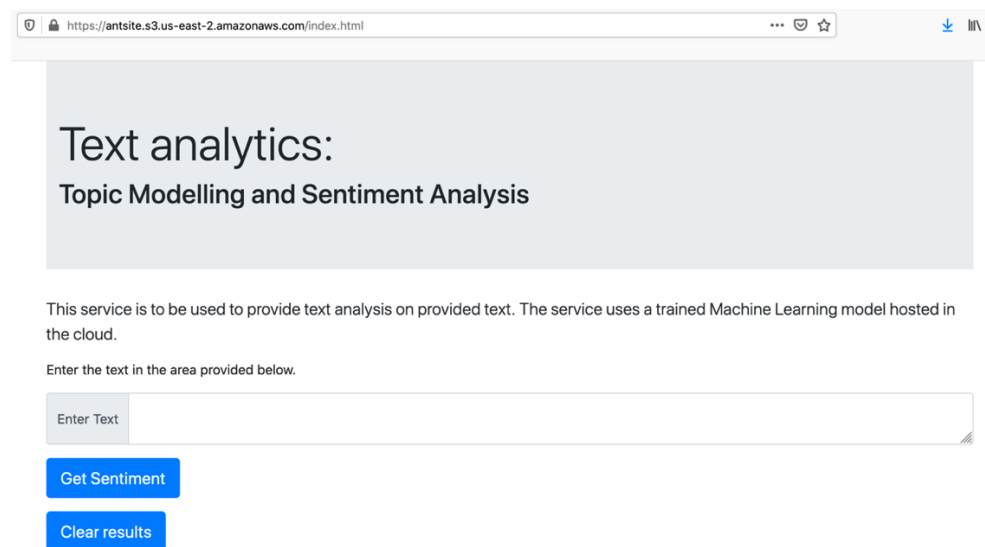


Figure 42 UI frontend of prototype

5.2.5 Dashboard development

The dashboard development involved the use of Amazon Quicksight, which is a serverless business intelligence service. The workflow involved in creating the dashboard is shown in *Figure 43*.

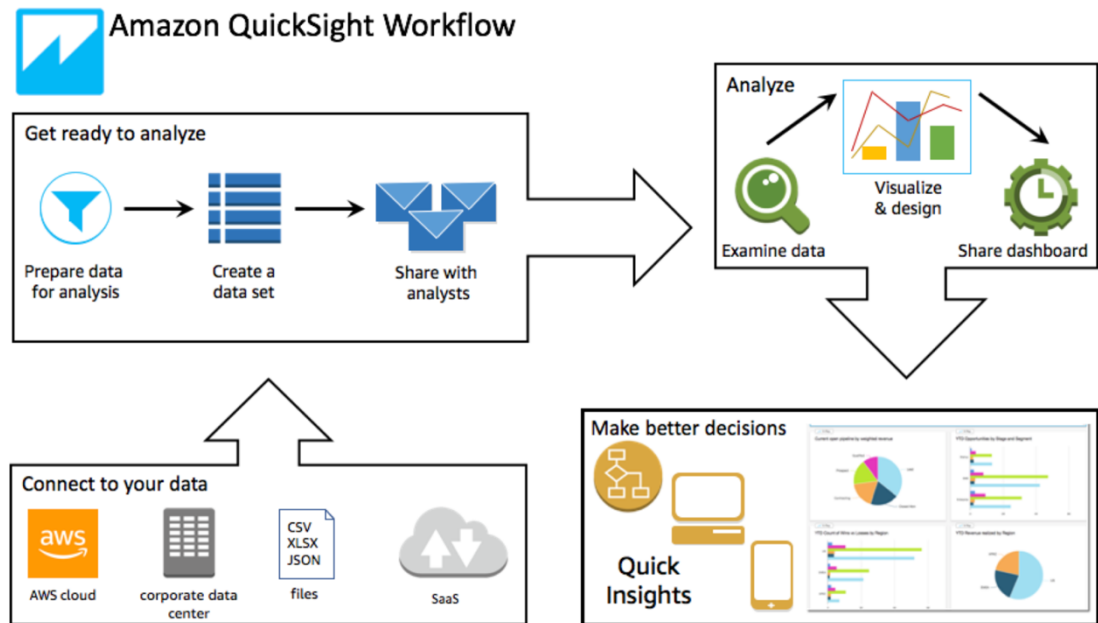


Figure 43 Quicksight Workflow [126]

Connecting the PostgreSQL database was made possible by inputting the details of the database as shown in Figure 44.

The screenshot shows the 'New PostgreSQL data source' configuration window. It contains the following fields and options:

- Data source name:** A text input field with a placeholder 'Enter a name for the data source'.
- Connection type:** A dropdown menu currently set to 'Public network'.
- Database server:** A text input field with a placeholder 'Enter hostname or IP address'.
- Port:** A text input field with a placeholder 'Port'.
- Database name:** A text input field with a placeholder 'Enter a database name'.
- Username:** A text input field with a placeholder 'Username'.
- Password:** A text input field with a placeholder 'Password'.
- Validation and Security:** At the bottom, there is a 'Validate connection' button, a checked checkbox for 'Enable SSL', and a blue 'Create data source' button.

Figure 44 Connect PostgreSQL to Quicksight

6 Testing

In this chapter, the tests run to evaluate the system are presented with the output received from the system. The purpose of running the test was to ensure the output of the system was expected. The output of the system was used to determine whether the prototype was working or whether further evaluations were required in line with industry recommendations [127].

6.1 Manual Testing

We decided to carry out manual testing of the system. To undertake the testing, the process steps used were to:

- i) enter the text into the textarea shown in the *Figure 42* frontend and clicking the “Get Sentiment” button
- ii) once the Sentiment Scores are received
- iii) click the “Get Topics” button
- iv) results of the Predicted Topics will be received, then preview the results.
- v) click “Clear results” button to clear the textarea and restart the process

6.1.1 Test 1:

Text: Vehicle tax Web site don't worry it hardly gets used anyway.....

Results:

Sentiment Scoring:

Sentiment Compound Score: 0.3412

Positive Score: 0.194

Negative Score: 0

Neutral Score: 0.806

Topics:

Topic1: 0.08 %

Topic2: 0.08 %
Topic3: 0.08 %
Topic4: 0.08 %
Topic5: 0.08 %
Topic6: 64.48 %
Topic7: 10.92 %
Topic8: 24.22 %

6.1.2 Test 2:

Text: passed my test...now give me my licence

Results:

Sentiment Scoring:

Sentiment Compound Score: 0

Positive Score: 0

Negative Score: 0

Neutral Score: 1

Topics:

Topic1: 0.14 %
Topic2: 19.92 %
Topic3: 19.92 %
Topic4: 19.92 %
Topic5: 19.92 %
Topic6: 19.92 %
Topic7: 0.14 %
Topic8: 0.14 %

6.1.3 Test 3:

Text: How infuriated would you be if the DVLA made an 'error', marked your driving licence as banned & your insurance was cancelled- will they fix it? Meanwhile an innocent guy has to stop everything & jump through hoops to 'fix' #DVLA error.. unbelievable! #getitsorted

Results:

Sentiment Scoring:

Sentiment Compound Score: -0.8658

Positive Score: 0.079

Negative Score: 0.262

Neutral Score: 0.659

Topics:

Topic1: 12.82 %

Topic2: 23.53 %

Topic3: 2.15 %

Topic4: 7.47 %

Topic5: 4.80 %

Topic6: 42.29 %

Topic7: 4.80 %

Topic8: 2.15 %

6.1.4 Test 4:

Text: I bet there's a lot too that are driving without actually passing their test. Paying someone to take their test for them
#dvla

Results:

Sentiment Scoring:

Sentiment Compound Score: 0

Positive Score: 0

Negative Score: 0

Neutral Score: 1

Topics:

Topic1: 7.21 %

Topic2: 7.21 %

Topic3: 16.61 %

Topic4: 7.21 %

Topic5: 16.61 %

Topic6: 45.08 %

Topic7: 0.04 %

Topic8: 0.04 %

6.1.5 Test 5:

Text: #DVLA



Figure 45 : Example Twitter Text "tweet"

Results:

Sentiment Scoring:

Sentiment Compound Score: 0

Positive Score: 0

Negative Score: 0

Neutral Score: 1

Topics:

Topic1: 12.5 %

Topic2: 12.5 %

Topic3: 12.5 %

Topic4: 12.5 %

Topic5: 12.5 %

Topic6: 12.5 %

Topic7: 12.5 %

Topic8: 12.5 %

6.1.6 Test 6:

Text: Hubby off work 5 months as DVLA suspended license after faint. No medical/ fitness to drive issues..all tests show nothing. Doctor said fit for work. Now 4 months waiting for #Dvla doc to sign a paper. Now have to claim #UniversalCredit as job was driving.Now mortgage at risk.

Results:**Sentiment Scoring:**

Sentiment Compound Score: -0.7514

Positive Score: 0.041

Negative Score: 0.184

Neutral Score: 0.775

Topics:

Topic1: 9.36 %

Topic2: 23.12 %

Topic3: 2.51 %

Topic4: 12.80 %

Topic5: 5.92 %

Topic6: 43.77 %

Topic7: 0.01 %

Topic8: 2.51 %

6.1.7 Test 7:

Text: Vehicle stopped in #Chester driver had his licence revoked on medical grounds by #DVLA & #GP & rightly so 2nd offence in 6 months. Vehicle seized & a long chat about #mentalhealth with a structured goal to getting back on track.

Results:**Sentiment Scoring:**

Sentiment Compound Score: -0.5233

Positive Score: 0

Negative Score: 0.117

Neutral Score: 0.883

Topics:

Topic1: 15.67 %

Topic2: 5.87 %

Topic3: 2.64 %

Topic4: 12.40 %

Topic5: 5.87 %

Topic6: 54.89 %

Topic7: 0.02 %

Topic8: 2.64 %

6.1.8 Test 8

Text: How long does it take for the application of a lost driving license? My mum applied as she's lost her paper one and has never had a photo ID license...she done this on 20th November and still hasn't had anything in the post?!

Results:

Sentiment Scoring:

Sentiment Compound Score: -0.5574

Positive Score: 0

Negative Score: 0.106

Neutral Score: 0.894

Topics:

Topic1: 0.03 %

Topic2: 19.96 %

Topic3: 0.03 %

Topic4: 7.61 %

Topic5: 3.52 %

Topic6: 65.31 %

Topic7: 0.03 %

6.1.9 Test 9

Text: @DVLAGovuk well this letter took a long time to get here, like 100 years!!! I must be living a second life or perhaps I've slipped through a #ParallelUniverse

Results:

Sentiment Scoring:

Sentiment Compound Score: 0.5574

Positive Score: 0.173

Negative Score: 0

Neutral Score: 0.827

Topics:

Topic1: 9.05 %

Topic2: 14.01 %

Topic3: 0.03 %
Topic4: 0.03 %
Topic5: 0.03 %
Topic6: 63.67 %
Topic7: 4.13 %
Topic8: 9.05 %

6.1.10 Test 10

Text: You have to love being able to tax your vehicle online with @DVLAgovuk. No more standing in post office queues and trying to find the right paperwork. It took me less than a minute! #cartax #DVLA #disability

Results:

Sentiment Scoring:

Sentiment Compound Score: 0.4588

Positive Score: 0.104

Negative Score: 0.054

Neutral Score: 0.842

Topics:

Topic1: 10.73 %

Topic2: 18.43 %

Topic3: 0.02 %

Topic4: 3.06 %

Topic5: 0.02 %

Topic6: 60.84 %

Topic7: 0.02 %

Topic8: 6.88 %

6.1.11 Test11

Text: When a car parked outside mine is reported nothing happens nothing ever happens so what's the point in paying my road tax #faithlostindvla

Results:**Sentiment Scoring:**

Sentiment Compound Score: 0

Positive Score: 0

Negative Score: 0

Neutral Score: 1

Topics:

Topic1: 10.43 %

Topic2: 10.43 %

Topic3: 10.43 %

Topic4: 0.04 %

Topic5: 0.04 %

Topic6: 58.17 %

Topic7: 0.04 %

Topic8: 10.43 %

6.2 Conclusion:

The results of the above tests are in line with the expectations of the system. The system is able to accept the entered text and request the sentiment scores and predictive topics. The response were as expected but further discussion such as the accuracy of the tests will be discussed further in the Discussion section.

7 Results and Evaluation

7.1 Test Results

Following the manual testing process highlighted in the Manual Testing section, the evaluation of the individual 8 topics of the deployed system was undertaken. *Table 6* shows a summary of the topics and the words that are contained in the individual topics.

Table 6 Relevant word(s) in the predictive topics

Name of Topic	Relevant word(s)
Topic 1	unlicensed
Topic 2	error, lost documents, lost licence
Topic 3	money
Topic 4	insurance, medical
Topic 5	test
Topic 6	application form, website, online enquiry, contact centre
Topic 7	police, enforcement
Topic 8	tax

The test coverage is shown in *Table 7*, and gives a reason why the tests were selected. Rather than give a specific mention of the topic categories returned or expected the sentiment, some of the information is to give a generalised view of topics.

Table 7 Test coverage

Test no.	Results			Reason for test	Valuation
	<i>System result</i>	<i>Expert expectation</i>	<i>Developer expectation (white box testing)</i>		Pass or Fail
1	Positive sentiment, 3 topics	Negative sentiment, 3 topics	Negative, 3 topics	Test the system works. The system returned the sentiment and topics	Fail
2	Failed to register the intention of the statement, gave positive sentiment rather than negative sentiment	The expert would expect the system to return negative sentiment	Expected system to return negative sentiment	Find out if the system can get semantics	Fail
3 & 4	Returned sentiment	To return relevant sentiment	Unsure as the Vader Sentiment was developed for short-form texts	Unsure whether the system can read multiple sentences?	Pass
5	Equal probability across the 8 topics	Expect the phrase used to cover all topics	Equal probability expected of 12.5%	Find out if a neutral text returns a neutral sentiment score and equal predictive tops	Pass
6-8	Returned a sentiment and topics covering	To return relevant sentiment	It should accept the text as it is from social media.	Longer text submitted from other social media site, Facebook.	Pass

9-11	Returned a positive sentiment	Expected a negative sentiment	Vader uses a cumulative scoring for sentiment, therefore expected it positive.	Can the system recognise sarcasm and deliver appropriate sentiment?	System does work as expected
------	-------------------------------	-------------------------------	--	---	------------------------------

The results show that use of Vader is not perfect as it falls short when handling some semantics. This is a known problem of lexicon-based systems.

7.2 Sentiment scores

To evaluate whether the system was providing the correct sentiment, an expert provided a sentiment score on the same texts used in the manual tests. The results are shown in *Table 8* below. The results show that the system provides a 64% accuracy of providing the correct sentiment. This result is better than a 50/50 guess and shows that the system although providing lower than the ideal result (where the expert is providing the sentiment).

Table 8 Comparison of Sentiment Results from the system and expert opinion

Test no.	Sentiment		Agree (1) Disagree (0)
	System	Expert opinion	
1	positive	negative	0
2	neutral	neutral	1
3	negative	negative	1
4	neutral	negative	0
5	neutral	neutral	1
6	negative	negative	1
7	negative	negative	1
8	negative	negative	1
9	positive	negative	0
10	positive	positive	1
11	neutral	positive	0
			<i>Agree = 7</i> <i>63.6%</i> <i>Disagree = 4</i> <i>36.4%</i>

Further evaluation of the system was done using sample texts to evaluate how the system compares to commercially available systems, expert and non-expert rating. We decided to determine the sentiment score in line with Vader Sentiment scoring [128] where a score of less than -0.05 is received the score is negative while a score of above 0.05 is a positive sentiment. Sentiment score of between -0.05 and 0.05 non-inclusive will be considered as a neutral sentiment.

Figure 46 shows that the system and the expert returned a neutral sentiment on the sample while the commercial and non-experts returned a negative sentiment. This shows that there is no significant difference between the system and using the expert. The commercial sentiment tool used for the comparison was the Microsoft Text Analytics API [129] and DataScience Toolkit [110].

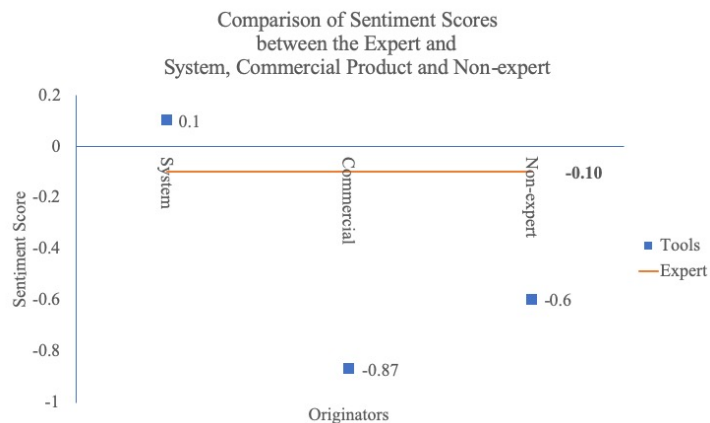


Figure 46 Comparison of sentiment scores

7.3 Topic modelling

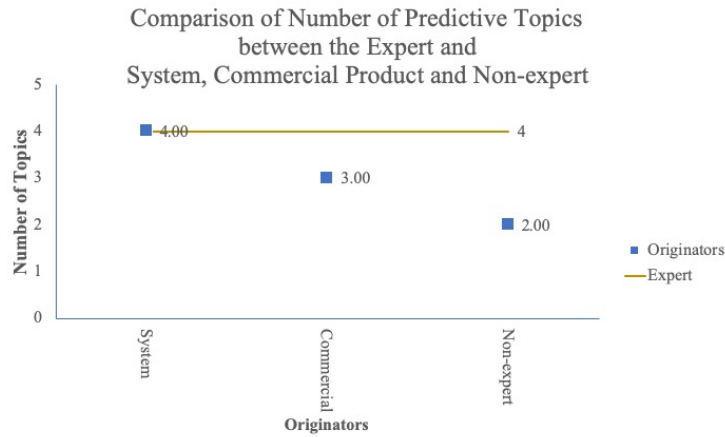


Figure 47 Comparison of predictive topics

Using the same sample used in *Figure 46*, *Figure 47* shows the average number of predictive topics. It shows that the system and expert returned a similar score on the topics while the commercial and non-expert returned a lower score on the topics. From this result, it shows that the system prediction for topic modelling is in line with the expert expectation. The commercial predictive topic modelling tool used for evaluation was BigML³ [130].

7.4 Growth of social media communication

Figure 48 shows the growth of Twitter communication as a form of communication between DVLA and its users. The chart shows the exponential growth that has corresponded with the operational change that has happened in the DVLA. Some of the major operational changes include the closure of the local office network in 2013 and the abolishment of the paper licence counterpart in 2015. It also follows the growth of services being delivered online in line with the Strategic Plans [131] [132].

³ BigML is a consumable, programmable, and scalable Machine Learning platform founded in 2011 and based in Corvallis, Oregon, and Valencia, Spain.

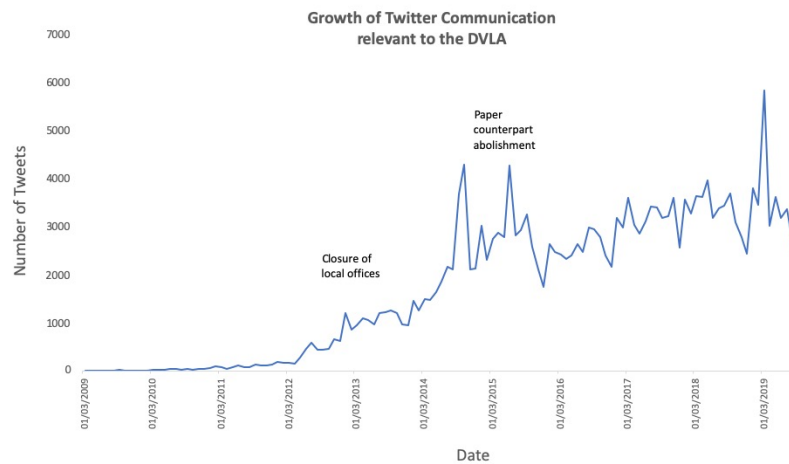


Figure 48 Growth of Twitter Communications relevant to DVLA

Figure 49 shows the breakdown of the topic probability over time, and it clearly shows an inverse relationship with the drop in topic 3 in years 2013 and 2015 and an increase for topic 4 and topic 6. This would follow with the expectation that with the closure of local offices and no paper licences more online interaction with the DVLA is about money, general DVLA service about applications and questions around drivers medical.

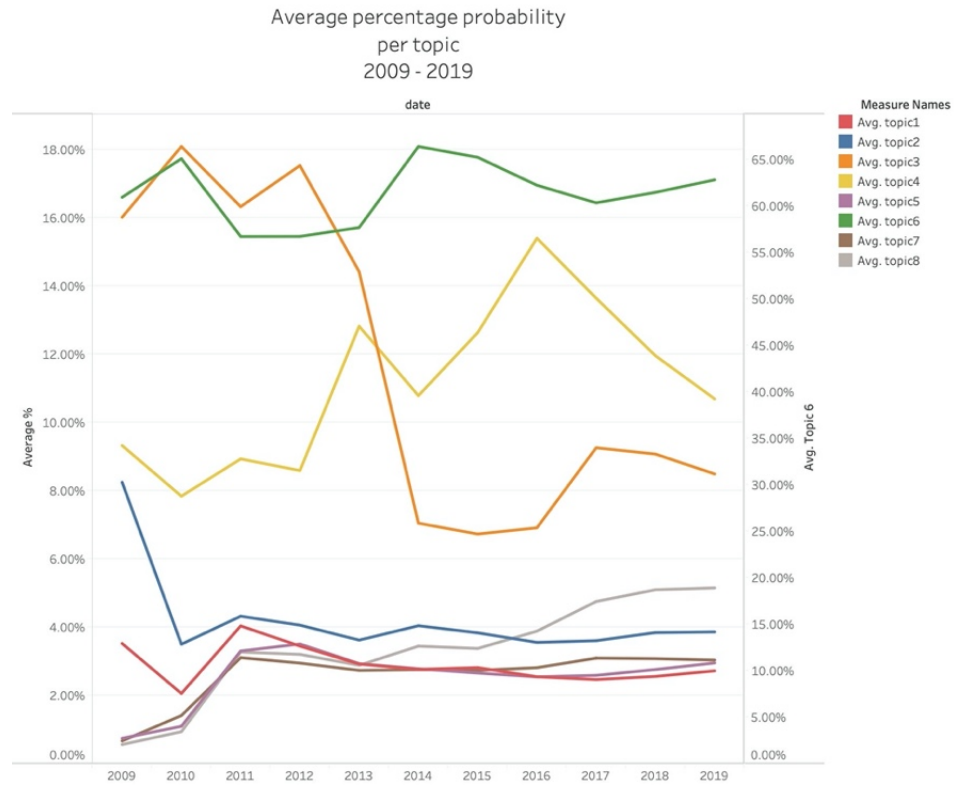


Figure 49 Average percentage probability per topic per year

7.5 Current state

From *Figure 49*, it is evident that the operational changes in 2013 and 2015 have led to changes within the topics, but with the changes, there are also implications on the sentiment. From *Figure 49*, we chose to focus on those topics that show changes namely Topic 3, Topic 4, Topic 6 and Topic 8. We then run the system using social media posts in January 2020.

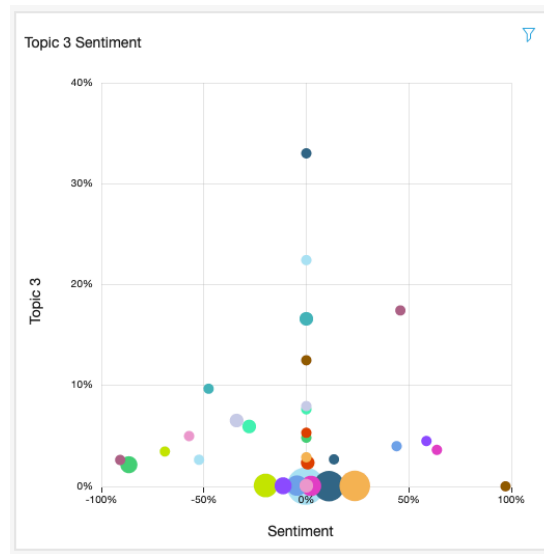


Figure 50 Topic 3 vs Sentiment

Figure 50 shows a cluster of sentiment around 0% and around 0% for topic prediction. The size of the bubble is representative to the count of run analysis. Taking those clusters out of the picture it shows that Topic 3 so far in January 2020 shows a strong negative sentiment. This is collaborated by the news reports about customers being targeted by scams by groups pretending to be the DVLA [128].

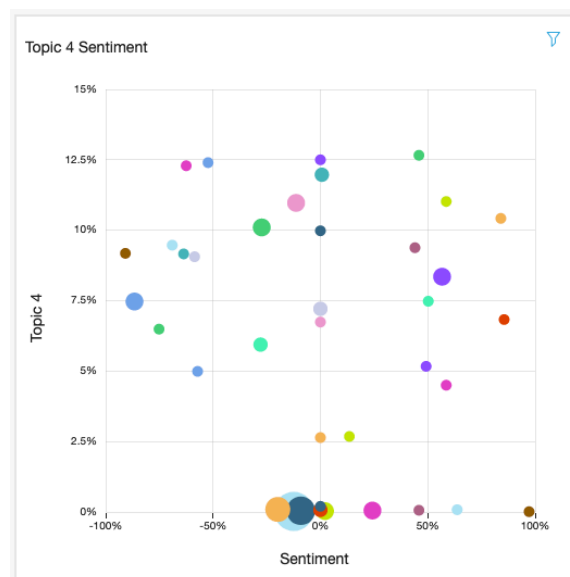


Figure 51 Topic 4 vs Sentiment

Figure 51 shows the sentiment around Topic 4. As with topic 3, a cluster has formed around 0 % topic prediction and around 0% sentiment score. Removing those results show that the sentiment for Topic 4 is balanced with an equal amount of positive and negative sentiment, although it is evident that there are instances of highly positive and highly negative sentiment. This is expected as the topic scope includes insurance, and medical issues.

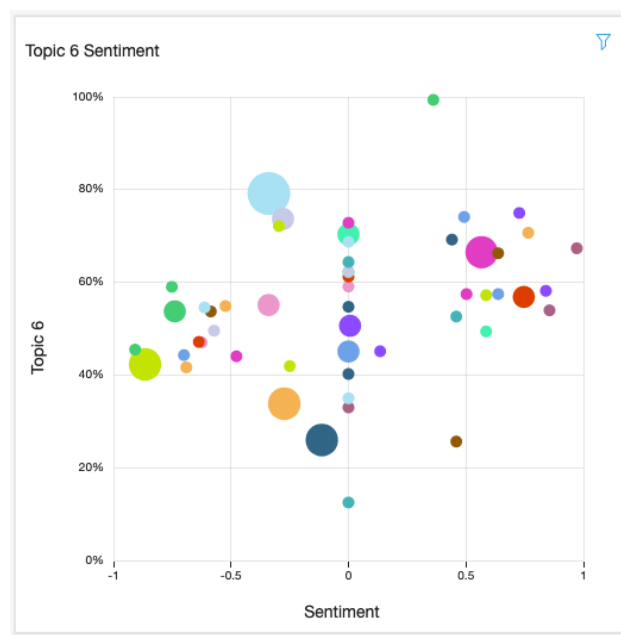


Figure 52 Topic 6 vs Sentiment

Topic 6 is represented in *Figure 52*, it shows that this topic is almost always part of the analysis as no 0% is present in the topic prediction. This topic scope there does include most of DVLA services. When we remove the cluster around 0% sentiment it shows that we have positive and negative clusters. However, considering the bubble sizes that represent count, it is clear that this topic is skewed to negative sentiment.

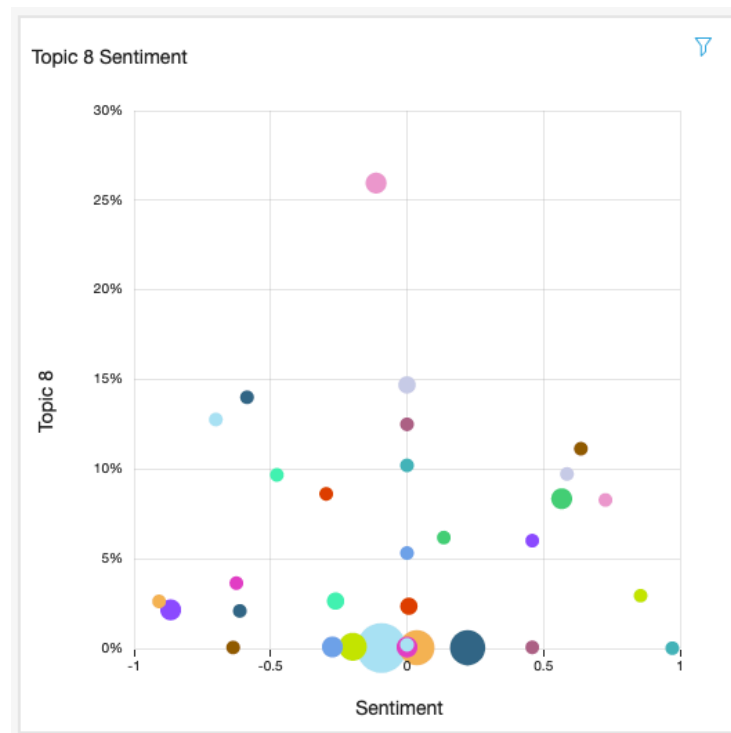


Figure 53 Topic 8 vs Sentiment

Topic 8 from *Figure 49*, shows slow growth in the influence of this topic. The topic scope includes tax. In *Figure 53*, when we remove the clusters around sentiment score 0% and topic prediction of 0%, the picture is a topic that has a negative sentiment score with a wide range in the topic probability prediction from single digits to 25%. It shows that issue around tax, which is now known as vehicle excise duty is raised in conjunction with other topics therefore the existence of the topic.

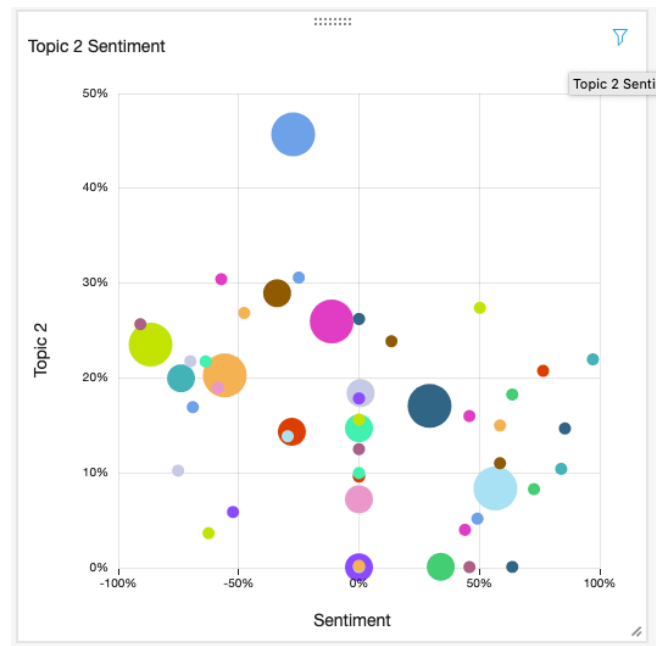


Figure 54 Topic 2 vs Sentiment

Topic 2 sentiment is shown in *Figure 54* and using *Table 6* is shown that Topic 2 reflects contact about errors, lost documents or lost licence. Notably, that the same issues were recently highlighted by the BBC consumer programme, BBC Watchdog [129]. The programme highlighted examples of DVLA customers being affected by errors in the service. *Figure 54* clearly shows that negative sentiment is the predominant expression for topic 2.

8 Discussion

From the results in section Results and Evaluation, we have been able to show that the use of predictive modelling and sentiment analysis makes it possible to identify trends occurring in the data. The prototype system results show that it was possible to find trends available in the customers' social media posts. For example, from the results, it has been possible to show that errors are one of the issues customers are contacting the organisation, *Figure 54*, and the sentiment is negative.

Sentiment analysis (SA) is an important area in customer service analysis. SA can be leveraged to assist in determining the level of customer experience, however, analysing unstructured text is difficult when it comes to the valence especially when sarcasm is involved. This is an area that needs improvement as the results show a 64% accuracy on sentiment scores on the given sample. To improve the sentiment analysis scores, it will be beneficial to improve the sentiment dictionary by adding domain and aspect specific lexicon making use of approaches such as pointwise mutual information approach (PMI) [135]. This will allow for adding terms such as “direct debit” not to be considered as a positive term just because the word direct is considered positive.

It is the nature of what social media is that most of the communication done through it is most likely to be a complaint rather than a compliment and as such the overall outlook of negative sentiment in by itself is not a surprise. It is, therefore, the onus on the organisations to compliment sentiment analysis with topic modelling. This will further allow for the identification of the topics of the communications. The efficacy of LDA has shown that it can uncover latent topics in social media posts. The use of LDA in the prototype system here is done using the Vowpal Wabbit wrapper which is implemented in the Microsoft Azure MLaaS service [136] [137].

Delivery of the prototype as a service deployed in the cloud would meet one of the goals of the strategic plan of the organisation [132] and in trend with what other government organisations are doing [138]. The use of MLaaS rather than developing systems from

scratch allows for the quick delivery of ideas from design to deployment by leveraging the existing services already being delivered through the cloud providers. Despite the advantages of running opensource machine learning frameworks in this case, NLTK, on owned hardware, issues of scaling and technical debt especially in scaling of the system would be resource-heavy. Therefore, the use of MLaaS provides an avenue of accessing advanced AI and data science skills packaged as a service, resource requirement are then handled as part of the service.

9 Conclusion

In conclusion, the advantages that sentiment analysis and predictive topic modelling offer a customer service team in an organisation is invaluable. The new world of customer service requires that it is not only important to know what the customer has said but it matters to also know how they feel. By having a system that can allow management to be aware of the issues as closest to real-time as possible, and with accuracy empowers the decision-makers to be able to make information-driven decisions.

The MLaaS services that many cloud providers are currently developing such as Microsoft Azure ML, Google GCP and AWS Machine Learning allow for organisations to make use of advanced AI, ML and data science skills without involving a lot more other than making use of the current existing contracts and the human resource (developer resource). Making use of the provided services would resolve some of the technical debt problems [10] existing in terms of analysis tools, serving infrastructure, monitoring and deployment. It is certainly the aim of the cloud providers that the services feel and look like the other services currently offered and therefore removing the need for their consumers to acquire a different skill set within the organisation.

The recommendation to the specific study organisation, DVLA, would be that the development of sentiment and predictive topic modelling service would be in line with the existing strategic plan of the organisation both in terms of developing services in the cloud and making use of new cloud technologies such as serverless. It would also enhance the reputation of the organisation in achieving its goal of being in the upper tier of customer service and experience.

Future work would involve more focus on the development of services through the use of MLaaS. We believe that the use of MLaaS will be an important factor in the future of business analytics and development. The demonstration of its uses in this project bears witness of the relevance of providing an effective simplification to the behemoth of big data and especially unstructured data of the form considered i.e. customers' social media interactions.

10 Reflection

Not a day goes by without some mention of artificial intelligence and machine learning in the news. At work, as part of the emerging technology team, it has given me a chance to discover topics that are of importance to the organisation. Customer service is part of the strategic plan of the DVLA and an area where the DVLA wants to ensure that it achieves “upper quartile in both digital and traditional channels”. The digital channel available in customer service is through the use of chatbots and social media such as Twitter, Instagram and Facebook.

My initial thoughts were to make use of the available chatbot interaction and find out how the “new” field in artificial intelligence and machine learning could be used to leveraged to understand the interactions between the DVLA and the customers. At present, the voice calls are transcribed to text and some analytics done on the resultant texts however the accuracy of the system has required the creation of a team to undertake manual interrogation of that data to confirm the output of the system. It felt to me that it should be possible to use AI and ML to make it a less human-intensive process.

However, what initially seemed a focussed idea ran into issues I had not anticipated. One, the subject involved a much steeper learning curve into what entails AI and ML. The field keeps receiving a lot of attention both in research, publication and profile. It was a struggle not having previously tackled the subject and for a long time initially, I spent most of the time reading on the subject matter and therefore my literature review period ended up taking longer than I had anticipated. I think if I had had teaching direction before the project, I would have been able to narrow my focus quicker and develop better and earlier. I was a novice in Python programming, which was central to the implementation of the project and therefore required time to upskill to allow for me to be able to proceed from design to development. However, the experience of learning Python was enjoyable due to the availability of training material and tutorials.

Surprisingly after the initial struggles of the knowledge acquiring and upskilling, once I got to the development section the process become easier. The use of cloud services in the use of AWS service and Microsoft services was easier to do as I could lean on previously acquired skills on cloud development at work. The implementation of the system was also helped by the fact that I had acquired skills through the course especially in the Webservice module, Agile module, and the Database module and as such nothing in the implementation of the system in terms of infrastructure was a challenge. During the development of the service in Azure, however, finding out the words relevant for the topics end up being a stab in the dark to try and find the relevant words of the categories and involved a manual testing exercise to discover the words. This is a particular gripe of users of the Azure MLS that Microsoft has not provided a tool to help the developer find out further detail of the makeup of the topics that are developed using the LDA.

For any future work, if I had more time, I would then proceed to the comparison through a review of MLaaS solutions and maybe compare the outcome of using commercial offerings against opensource platforms such as using Apache Platform AI. I would also expand the system to allow for file upload facility for analysis. I would also add a streaming facility possibility using the streaming technologies such as Spark to develop a way of getting the social media posts in real-time rather than through the current system where the posts have to be copied across. This would allow for less human intervention and free the human resource time involved.

Finally, looking back I most of all wish I had more knowledge earlier in the process and had more time available. What I take from this experience is that I have discovered a fascinating field and it is so wide and I will continue to investigate and strive to continue to increase my knowledge and skills.

The whole M.Sc. experience at the DVLA, and UWTSD has been a blessing for me and my family and I count myself honoured by the experience. Thanks.

11 References

- [1] R. Blumberg and S. Atre, "The Problem with Unstructured Data," *DM Review*, vol. Feb, no. Feb, pp. 42-46, 2003.
- [2] V. Dhar, "Data science and prediction," *Communications of the ACM*, vol. 56, no. 12, pp. 64-73, 2013.
- [3] D. Davis, "AI Unleashes the Power of Unstructured Data," *CIO*, 9 July 2019.
- [4] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, no. 2, pp. 137-144, 2015.
- [5] Enterprise Software Team, Merrill Lynch, "Enterprise Information Portals," Merrill Lynch, New York, 1998.
- [6] B. Polnaszek, A. Gilmore-Bykovskyi and A. J. Kind, "Overcoming the Challenges of Unstructured Data in Multi-site, Electronic Medical Record-based Abstraction," *Med Care*, vol. 54, no. 10, pp. 65-72, 2016.
- [7] S. VanBoskirk, S. C. Overby and S. Takvorian, "US Interactive Marketing Forecast, 2011 To 2016," Forrester, Cambridge, MA, 2011.
- [8] C. Sashi, "Customer engagement, buyer-seller relationships, and social media," *Management Decision*, vol. 50, no. 2, pp. 253-272, 2012.
- [9] B. Evelson, S. Sridharan and R. Perdoni, "Market Overview: Text Analytics," Forrester Research, Cambridge, MA, 2015.
- [10] D. e. Sculley, "Hidden Technical Debt in Machine Learning Systems," in *Advances in Neural Information Processing Systems*, Montreal, 2015.
- [11] DVLA, "DVLA," DVLA, 01 09 2019. [Online]. Available: <https://www.gov.uk/government/organisations/driver-and-vehicle-licensing-agency>. [Accessed 01 09 2019].
- [12] Gartner Research, "Market Guide for Text Analytics," Gartner Research, Stamford, 2015.
- [13] W. Chung, "BizPro: Extracting and categorizing business intelligence factors from textual news articles," *International Journal of Information Management*, vol. 34, no. 2, pp. 272-284, 2014.
- [14] Markets and Markets, "Text Analytics Market by Component (Software, Services), Application (Customer Experience Management, Marketing Management, Governance, Risk and Compliance Management), Deployment Model, Organization Size, Industry Vertical, Region - Global Forecast," Markets and Markets, Northbrook, 2019.
- [15] SAS, "Predictive Analytics," SAS, Cary, NC, 2019.
- [16] R. Lowe, N. Pow, I. Serban and J. Pineau, "The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems," in

16th Annual SIGdial Meeting on Discourse and Dialogue, SIGDIAL 2015, Prague, Czech Republic, 2015.

- [17] P. McCorduck, *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*, Baco Raton: A K Peters / CRC Press, 2004.
- [18] J. McCarthy, "What is Artificial Intelligence," *John McCarthy Articles, Stanford University*, pp. 2-15, 12 November 2007.
- [19] S. L. Tanimoto, *The elements of artificial intelligence using Common Lisp*, New York: Computer Science Press, 1990.
- [20] S. J. Russell and P. Norvig, "Artificial Intelligence : A Modern Approach," New Jersey, Prentice Hall, 1995, pp. 1-29.
- [21] E. Charniak and D. McDermott, *Introduction to Artificial Intelligence*, Boston, Mass.: Addison-Wesley Publishing Company, 1985.
- [22] P. H. Winston, *Artificial Intelligence*, Boston, Mass.: Addison-Wesley Publishing Compan, 1992.
- [23] R. J. Schalkoff, *Artificial Intelligence: An Engineering Approach*, New York: McGraw-Hill College, 1990.
- [24] G. F. Luger and W. A. Stubblefield, *Artificial intelligence: structures and strategies for complex problem solving*, San Francisco: The Benjamin/Cummings Pub. Co, 1993.
- [25] A. Turing, "COMPUTING MACHINERY AND INTELLIGENCE," *Mind*, vol. LIX, no. 236, p. 433–46, 1950.
- [26] M. A. Boden, "Creativity and artificial intelligence," *Artificial Intelligence*, vol. 103, no. 1-2, pp. 347-356, 1998.
- [27] J. J. Staszewski, "Cognitive Models of Human Expertise and their Scientific and Practical Value," in *Association for the Advancement of Artificial Intelligence*, Arlington, Virginia, 2008.
- [28] A. Varga-Szemes, B. E. Jacobs and J. Schoepf, "The power and limitations of machine learning and artificial intelligence in cardiac CT," *Journal of Cardiovascular Computed Tomography*, vol. 12, no. 3, pp. 202-203, 2018.
- [29] E. Siegel, *Predictive Analytics*, Hoboken, New Jersey: John Wiley and Sons, Inc, 2013.
- [30] M. Minsky, "Steps Toward Artificial Intelligence," *PROCEEDINGS OF THE IRE*, pp. 8-30, January 1960.
- [31] D. Barber, *Bayesian Reasoning and Machine Learning*, Cambridge: Cambridge University Press, 2012.
- [32] A. Newell and H. A. Simon, *Human Problem Solving*, Englewood Cliffs, NJ: Prentice-Hall, 1972.
- [33] H. A. Simon and A. Newell, "Heuristic problem solving: The next advance in operations research," *Operation Research*, vol. 6, no. 1, Jan-Feb 1958.

- [34] O. Selfridge, "Self Pattern recognition and modern computers," in *Proceedings of the Western Joint Computer Conference*, 1955.
- [35] R. Fagin, J. Y. Halpern, Y. Moses and M. Y. Vardi, Reasoning about Knowledge, Cambridge, Massachusetts; London, England: The MIT Press, 1995.
- [36] N. J. Nilsson, "Shakey the Robot - Technical Note 323," SRI International, Menio Park, CA, 1984.
- [37] W. N. Locke and A. D. Booth, "Machine Translation of Languages: Fourteen Essays," Technology Press, Cambridge, MA, 1955.
- [38] J. Weizenbaum, "ELIZA - a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36-45, 1966.
- [39] T. Winograd, "Procedures as a Representation for Data in a Computer Program for Understanding Natural Language," MIT, Cambridge, 1971.
- [40] T. Bickmore and et.al, "Patient and Consumer Risks when using conversational assistants for medical information: an observational study of Siri, Alexa and Google Assistant," *Journal of Medical Internet Research*, vol. 20, no. 9, 2018.
- [41] A. L. Samuel, "Some Studies in Machine Learning using the Game of Checkers," *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210-229, 1959.
- [42] T. M. Mitchel, Machine Learning, New York: McGraw-Hill Education, 1997.
- [43] K. P. Murphy, Machine Learning: A Probabilistic Perspective, Cambridge, Massachusetts; London, England: MIT Press, 2012.
- [44] J. R. Quinlan, "Induction of Decision Trees," *Machine Learnin*, vol. 1, no. 1, pp. 81-106, 1986.
- [45] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," *Proceedings of the Fifth Berkeley Symposium on Mathematics and Probability*, vol. Volume 1, pp. 281-297, 1967.
- [46] J. Leskovec, A. Rajaraman and J. D. Ullman, "Clustering," in *Mining of Massive Datasets*, New York, NY, Cambridge University Press, 2011, pp. 257-258.
- [47] K. F. Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, vol. 2:11, pp. 559-572, 1901.
- [48] H. Hotelling, "Relations Between Two Sets of Variates," *Biometrika Trust*, vol. 28, no. 3/4, pp. 321-277, 1936.
- [49] W. Maass and T. Gyorgy, "Lower Bound Methods and Separation Results for On-Line Learning Models," *Machine Learning*, vol. 9, no. 2-3, pp. 107-145, 1992.
- [50] D. Helmbold, R. Sloan and M. K. Warmuth, "Learning Nested Differences of Intersection-Closed Concept Classes," *Machine Learning*, vol. 5, no. 2, pp. 165-196, 1990.

- [51] L. B. Booker, “Classifier systems that learn internal world models,” *Machine Learning*, vol. 3, no. 2-3, pp. 161-192, 1988.
- [52] D. W. Aha, D. Kibler and M. K. Albert, “Instance-based Learning Algorithms,” *Machine Learning*, vol. 6, no. 1, pp. 37-66, 1991.
- [53] T. G. Dietterich and R. S. Michalski, “Inductive learning of structural descriptions: Evaluation criteria and comparative review of selected methods,” *Artificial Intelligence*, vol. 16, no. 3, pp. 257-294, 1981.
- [54] A. Geron, *The Machine Learning Landscape*, Sebastopol, California: O'Reilly Media, 2017, pp. 3-32.
- [55] H. Kang, S. Joon Yoo and D. Han, “Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews,” *Expert Syst Appl.*, vol. 39, pp. 6000-6010, 2012.
- [56] O. Chapelle, B. Scholkopf and A. Zien, *Semi-Supervised Learning*, Cambridge, Massachusetts; London, England: The MIT Press, 2005.
- [57] R. E. Bellman, *Dynamic programming*, Princeton, NJ: Princeton University Press, 1957.
- [58] C. Sammut (Ed) and G. i. Webb (Ed), “Curse of Dimensionality,” in *Encyclopedia of Machine Learning*, New York, NY, Springer US, 2017, pp. 257-258 .
- [59] R. S. Sutton and A. G. Barto, *Reinforcement Learning: an introduction*, Cambridge, MA; London, England: MIT Press, 2018.
- [60] R. Howard, *Dynamic Programming. and Markov Process*, Cambridge, MA: MIT Press, 1960.
- [61] R. Bellman, *Dynamic Programming*, Princeton, NJ: Princeton University Press, 1957.
- [62] D. e. a. Silver, “Mastering the game of Go without human knowledge,” *Nature*, vol. 550, pp. 354-359, 2017.
- [63] M. Hildebrandt, “Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning,” *Theoretical Inquiries in Law*, vol. 20, no. 1, pp. 83-121, 2019.
- [64] S. Barocas and A. D. Selbst, “Big Data’s Disparate Impact,” *California Law Review*, vol. 104, no. 3, pp. 672-729, 2016.
- [65] G. C. Platform, Composer, *Machine Learning Bias and Fairness with Timnit Gebru and Margaret Mitchell: GCPPodcast 114*. [Sound Recording]. Google. 2018.
- [66] L. Dixon, J. Li, J. Sorensen, N. Thain and L. Vasserman, “Measuring and Mitigating Unintended Bias in Text Classification,” in *AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, New Orleans, 67–73.

- [67] R. K. Bellamy, K. Day, M. Hind and et.al, "AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias," *ArXiv*, vol. arXiv:1810.01943, 2018.
- [68] J. A. Adebayo, "FairML: Auditing Black-Box Predictive Models," Github, 23 March 2017. [Online]. Available: <https://github.com/adebayoj/fairml>. [Accessed 13 September 2019].
- [69] R. Feldman and J. Sanger, *The Text Mining Handbook*, Cambridge, MA: Cambridge University Press, 2007.
- [70] M. Hearst, "What Is Text Mining?," University of California, Berkeley, California, 2003.
- [71] D. Swanson, "Two medical literatures that are logically but not bibliographically connected," *Journal of the American Society for Information Science*, vol. 38, no. 4, pp. 228-233, 1987.
- [72] L. Guernsey, "Digging For Nuggets Of Wisdom," *New York Times*, New York, 2003.
- [73] C. C. Aggarwal, *MAchine learning for text*, Cham: Springer, 2018.
- [74] E. D. Liddy, "Natural Language Processing," in *In Encyclopedia of Library and Information Science*, New York, Marcel Decker, Inc. , 2001.
- [75] W. P. PR, "Washington Post," *Washington Post*, 5 August 2016. [Online]. Available: <https://www.washingtonpost.com/pr/wp/2016/08/05/the-washington-post-experiments-with-automated-storytelling-to-help-power-2016-rio-olympics-coverage/>. [Accessed 2 September 2019].
- [76] J. Whitwell and C. Molumby, "General election 2019: Semi-automation makes it a night of 689 stories," *BBC NEWS LABS*, London, 2019.
- [77] W. Medhat, A. Hassan and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093-1113, 2014.
- [78] R. Sharma, M. Gupta, A. Agarwal and P. Bhattacharyya, "Adjective Intensity and Sentiment Analysis," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2015.
- [79] R. Quinlan, "Induction of decision trees," *Machine Learn*, vol. 1, pp. 81-106, 1986.
- [80] F. Xianghua, L. Guo, G. Yanyan and W. Zhiqiang, "Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon," *Knowledge based system*, vol. 37, pp. 186-195, 2013.
- [81] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.
- [82] L. Sperkova, "Review of Latent Dirichlet Allocation Methods Usable in Voice of Customer Analysis," *ACTA INFORMATICA PRAGENSIA*, vol. 7, no. 2, pp. 152 - 165, 2018.

- [83] Raconteur, "Future Data 2019," Raconteur, London, 2019.
- [84] K. Adnan, R. Akbar, S. W. Khor and A. B. A. Ali, "Role and Challenges of Unstructured Big Data in Healthcare," in *Advances in Intelligent Systems and Computing*, Singapore, 2019.
- [85] D. Reinsel, J. Gantz and J. Rydning, "The Digitization of the World From Edge to Core," IDC, Framingham, MA, 2018.
- [86] IBM, "Cloud computing: A complete guide," IBM, [Online]. Available: <https://www.ibm.com/cloud/learn/cloud-computing>. [Accessed 01 09 2019].
- [87] M. Metheny, *Federal Cloud Computing - The Definitive Guide for Cloud Service Providers*, Cambridge, MA: Syngress, 2017.
- [88] M. Veverka, "A Private Party," Barron's, New York, NY, 2010.
- [89] A. Khayer, S. M. Talukder, Y. Bao and N. M. Hossain, "Cloud computing adoption and its impact on SMEs' performance for cloud supported operations: A dual-stage analytical approach," *Technology in Society*, vol. 20, 2019.
- [90] M. Nunnikhoven, "Putting AWS Security Services to Work For You," 14 March 2019. [Online]. Available: <https://aws.amazon.com/blogs/aws/aws-heroes-putting-aws-security-services-to-work-for-you/>. [Accessed 01 09 2019].
- [91] M. R. Ferre, "Compute Abstractions on AWS: A Visual Story," 06 09 2018. [Online]. Available: <https://aws.amazon.com/blogs/architecture/compute-abstractions-on-aws-a-visual-story/>. [Accessed 01 09 2019].
- [92] AWS, "Building Applications with Serverless Architectures - AWS Serverless Multi-Tier Architectures," AWS, Seattle, 2015.
- [93] AWS, "Amazon Relational Database Service (RDS)," AWS, 01 09 2019. [Online]. Available: <https://aws.amazon.com/rds/>. [Accessed 01 09 2019].
- [94] insideBigData, "Machine Learning as a Service (MLaaS) and its business implications in 2019," insideBigData, Portland, OR, 2019.
- [95] Accenture, "Accenture Touchless Testing Platform - Real Time Analytics Platform," Accenture, New York, NY, 2018.
- [96] ADWEEK, "Deloitte Is Buying a Marketing AI Platform to Expand Its Ad-Tech Holdings," ADWEEK, New York, NY, 2018.
- [97] AWS, "Amazon Quicksight," AWS, Seattle, 2019.
- [98] Kaggle, "Kaggle," Kaggle, 01 09 2019. [Online]. Available: <https://www.kaggle.com/datasets>. [Accessed 01 09 2019].
- [99] KD Nuggets, "KD Nuggets," KD Nuggets, 01 09 2019. [Online]. Available: <https://www.kdnuggets.com/datasets/index.html>. [Accessed 01 09 2019].
- [100] AWESOME Data, "Awesome Public Datasets," Github, 01 09 2019. [Online]. Available: <https://github.com/awesomedata/awesome-public-datasets>. [Accessed 01 09 2019].
- [101] Gov UK, "Data Gov UK," Gov UK, 01 09 2019. [Online]. Available: <https://data.gov.uk/>. [Accessed 01 09 2019].

- [102] Microsoft, "Microsoft Research Open Data," Microsoft, 01 09 2019. [Online]. Available: <https://msropendata.com/>. [Accessed 01 09 2019].
- [103] Google, "Google Dataset Search," Google, 01 09 2019. [Online]. Available: <https://toolbox.google.com/datasetsearch>. [Accessed 01 09 2019].
- [104] D. Yan and G. E. Davis, "A First Course in Data Science," *Journal of Statistics Education*, vol. 27, no. 2, pp. 99-109, 2019.
- [105] D. Conway, "THE DATA SCIENCE VENN DIAGRAM," Drew Conway, 30 09 2010. [Online]. Available: <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>. [Accessed 01 09 2019].
- [106] Python.org, "Python," Python.org, 01 09 2019. [Online]. Available: <https://www.python.org/>. [Accessed 01 09 2019].
- [107] Anaconda, "Anaconda," Anaconda, 01 09 2019. [Online]. Available: <https://www.anaconda.com/distribution/>. [Accessed 01 09 2019].
- [108] Jupyter, "The Jupyter Notebook," Jupyter Notebook, 01 09 2019. [Online]. Available: <https://jupyter.org/>. [Accessed 01 09 2019].
- [109] S. Bird, E. Klein and E. Loper, *Natural Language Processing with Python - Analyzing Text with the Natural Language Toolkit*, Sebastopol, CA: O'Reilly Media, 2009.
- [110] P. Warden, "Data Science Toolkit," Data Science Toolkit, 2013. [Online]. Available: <http://www.datasciencetoolkit.org/>. [Accessed 01 09 2019].
- [111] A. Stellman and J. Greene, *Applied Software Project Management*, Sebastopol, CA: O'Reilly Media, 2005.
- [112] Z. Michalewicz, M. Schmidt, M. Michalewicz and C. Chiriac, "Adaptive Business Intelligence: Three Case Studies," in *Evolutionary Computation in Dynamic and Uncertain Environments*, Berlin, Springer-Verlag , 2007, pp. 179-196.
- [113] CS McGill University, "Ubuntu Dialogue Corpus v1.0.," CS McGill University, 2015. [Online]. Available: <http://dataset.cs.mcgill.ca/ubuntu-corpus-1.0/>. [Accessed 01 09 2019].
- [114] DVLA, "DVLA Twitter," DVLA, 21 09 2019. [Online]. Available: <https://twitter.com/DVLAgovuk>. [Accessed 21 09 2019].
- [115] Jonbakerfish, "GITHUB," GITHUB, 12 09 2019. [Online]. Available: <https://github.com/jonbakerfish/TweetScraper>. [Accessed 01 10 2019].
- [116] J. Singh and V. Gupta, "Text Stemming: Approaches, Applications, and Challenges," *ACM Computing Surveys (CSUR)*, vol. 49, no. 3, p. Article 45, 2016.
- [117] matplotlib, "Matplotlib v 3.1.2," matplotlib, 01 09 2019. [Online]. Available: <https://matplotlib.org/>. [Accessed 01 09 2019].
- [118] PyPI, "pyLDavis," PyPi, 01 09 2019. [Online]. Available: <https://pypi.org/project/pyLDavis/>. [Accessed 01 09 2019].

- [119] C. Mascolo, "Software Design Models, Tools & Processes," 2019. [Online]. Available: <https://www.cl.cam.ac.uk/teaching/1112/SWDesign/softwaredesign01.pdf>. [Accessed 01 11 2019].
- [120] C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, Ann Arbor, MI, 2014.
- [121] A. K. McCallum, "MALLET: A Machine Learning for Language Toolkit," University of Massachusetts Amherst, Amherst, Massachusetts, 2002.
- [122] R. Rehurek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta, 2010.
- [123] PyPi, "Gensim - Python framework for fast Vector Space Modelling," PyPI, 09 26 2019. [Online]. Available: <https://pypi.org/project/gensim/>. [Accessed 01 10 2019].
- [124] Gensim, "Latent Dirichlet Allocation via Mallet," Gensim, 2019. [Online]. Available: <https://radimrehurek.com/gensim/models/wrappers/ldamallet.html>. [Accessed 01 10 2019].
- [125] M. Roder, A. Both and A. Hinneburg, "Exploring the Space of Topic Coherence Measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, New York, NY, 2015.
- [126] AWS, "Amazon Quicksight Userguide," 2019. [Online]. Available: <https://docs.aws.amazon.com/quicksight/latest/user/welcome.html>. [Accessed 01 09 2019].
- [127] GOV.UK , "GOV.UK Service Manual," GOV.UK , 10 2016. [Online]. Available: <https://www.gov.uk/service-manual/design/making-prototypes>. [Accessed 01 09 2019].
- [128] C. Hutto, "Vader Sentiment Github," Github, 2017. [Online]. Available: <https://github.com/cjhutto/vaderSentiment#about-the-scoring>. [Accessed 01 10 2019].
- [129] Microsoft, "Microsoft Azure Text Analytics," Microsoft, 01 2020. [Online]. Available: <https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/>. [Accessed 01 2020].
- [130] BigML, "BigML," BigML, 01 2020. [Online]. Available: <https://bigml.com/>. [Accessed 01 2020].
- [131] DVLA, "DVLA Strategic Plan for 2014 to 2017," DVLA, Swansea, 2017.
- [132] DVLA, "DVLA Strategic Plan for 2017 to 2020," DVLA, Swansea, 2018.
- [133] Moneywise, "DVLA kicks off 2020 with scam warnings," Moneywise, London, 2020.
- [134] BBC, "DVLA Data Breaches Update," BBC Watchdog, Salford, UK, 2019.

- [135] J. Bross and H. Ehrig, "Automatic construction of domain and aspect specific sentiment lexicons for customer review mining," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, San Francisco, CA, 2013.
- [136] M. Hoffman, F. R. Bach and D. M. Blei, "Online Learning for Latent Dirichlet Allocation," in *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, Vancouver, BC, 2010.
- [137] Microsoft, "Latent Dirichlet Allocation," 05 06 2019. [Online]. Available: <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/latent-dirichlet-allocation>. [Accessed 01 09 2019].
- [138] Gartner, "Understanding Cloud Adoption in Government," Gartner, Stamford, 2018.
- [139] IBM, "AI Fairness 360," IBM, 19 September 2018. [Online]. Available: <http://aif360.mybluemix.net/>. [Accessed 6 September 2019].
- [140] T. Calders and S. Verwer, "Three naive Bayes approaches for discrimination-free classification," *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 277-292, 2010.
- [141] S. Marsland, *Machine Learning: An Algorithmic Perspective*, Boca Raton: CRC Press, 2015.
- [142] C. Darwin, *On the Origin of Species by Means of Natural Selection*, London: John Murray, 1859.
- [143] B. Xue, M. Zhang, W. N. Browne and X. Yao, "A Survey on Evolutionary Computation Approaches to Feature. Selection," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 4, pp. 606-626, 2016.
- [144] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*, Cambridge, Massachusetts: MIT Press, 2016.
- [145] L. M. Minsky and S. Papert, "Perceptrons: An Introduction to Computational Geometry," *IEEE Transactions on Information Theory*, vol. 15, no. 6, pp. 738-739, 1969.
- [146] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115-133, 1943.
- [147] H. Webb, *The Organization of Behaviour: A NeuroPhyschological Theory*, New York, NY; London, England: John Wiley & Sons; Chapman & Hall, 1949.
- [148] K. Fukushima, "Neocognitron: A Self-organizing Neural Network Model for a Mechanisim of Pattern Recognition Unaffected by Shift in Position," *Biological Cybernetics*, vol. 36, pp. 193-202, 1980.
- [149] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, Lake Tahoe, 2012.

- [150] K. He, X. Zhang, S. Ren and J. Sun, “Deep Residual Learning for Image Recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, 2016.
- [151] Gartner Research, “Gartner Glossary,” Garner, 10 Sept 2019. [Online]. Available: <https://www.gartner.com/en/information-technology/glossary/big-data>. [Accessed 10 Sept 2019].

12 Appendix

12.1 SRS

12.1.1 Use Cases:

Analyst

UC-2: As an analyst I want to be able to run the sentiment analysis.

UC-3: As an analyst I want to be able to run the topic modelling analysis

UC-4: As an analyst I should be able to save the results for the sentiment analysis and topic identification for the entered text for later analysis

Manager:

UC- 5 As a manager I want to be able to visually see the daily analysis of system.

12.2 Design and Development

12.2.1 Pre-processing

12.2.1.1 Ubuntu pre-processing

```
import pandas as pd

def f(x):
    op = pd.Series(dict(date="{%s}" % ' '.join(x['date']),
                      fromWho="{%s}" % ' '.join(x['from']),
                      to="{%s}" % ' '.join(x['to']),
                      text="{%s}" % ' '.join(x['text'])))
    op.at['date'] = op.get('date')[1:-1].split(' ')[0]
    lo = op.get('fromWho')[1:-1].split(' ')
    op.at['fromWho'] = lo[0]
    receiver = None
    for sender in lo:
        if sender!=lo[0]:
            receiver = sender
            break
    op.at['to'] = receiver
```

```

        return op

# import the file
df = pd.read_csv("dialogueText_301.csv")
df = df.applymap(str)
print('Getting the Dataframe to String')

# merging the chat session to one row per conversation
fd = df.groupby("dialogueID").apply(f)
print('Writing the Dataframe to csv')

# print(pd.DataFrame.to_string(fd))
fd.to_csv('Merge.csv')

```

12.2.1.2 DVLA Tweets pre-processing

```

import os, json, csv
import pandas as pd

# this finds our json files
path_to_json = 'tweet/'
json_files = [pos_json for pos_json in os.listdir(path_to_json)]

#Print to inform the number of JSON files
print('Number of files is:' , len(json_files))

# This is to define the pandas Dataframe with the columns I want to extract
from the captured tweet JSON
jsons_data = pd.DataFrame(columns=['usernameTweet','ID', 'text',
'nbr_reply','is_reply', 'datetime','is_retweet'])

# we need both the json and an index number so use enumerate()
for index, js in enumerate(json_files):
    with open(os.path.join(path_to_json, js)) as json_file:
        json_text = json.load(json_file)
        # Layout of the Tweet JSON
        username = json_text['usernameTweet']
        ids = json_text['ID']
        text = json_text['text']
        nbrreply = json_text['nbr_reply']
        reply = json_text['is_reply']
        datetime = json_text['datetime']
        retweet = json_text['is_retweet']

```

```

        # this is to push the data into a pandas DataFrame at the row given by
        'index'
        jsons_data.loc[index] = [username, ids, text,
        nbrreply,reply,datetime,retweet]

# Finally, write the DataFrame to a csv file
jsons_data.to_csv('dvl tweets.csv', encoding='utf-8', index=False)
#information only notification on completion
print('Writing Complete')

```

12.2.1.3 Topic Modelling pyLDAvis Output

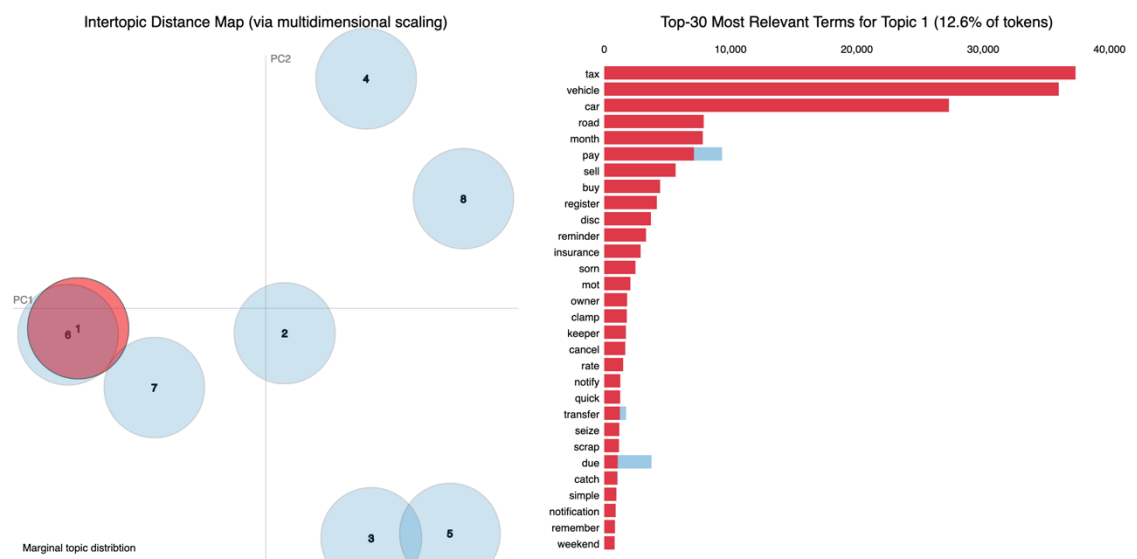


Figure 55 LDA Topic 1

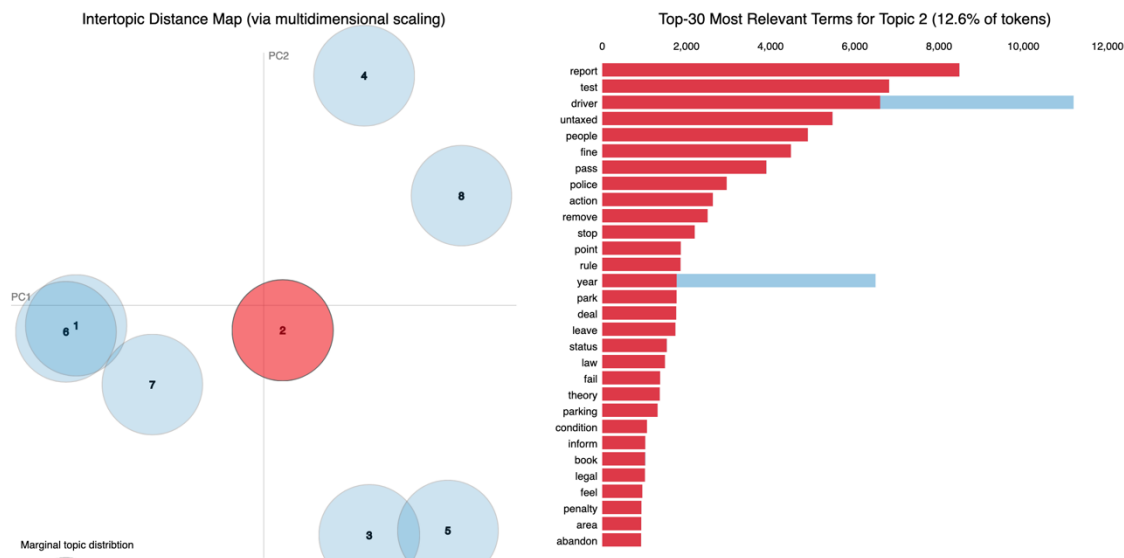


Figure 56 LDA Topic 2

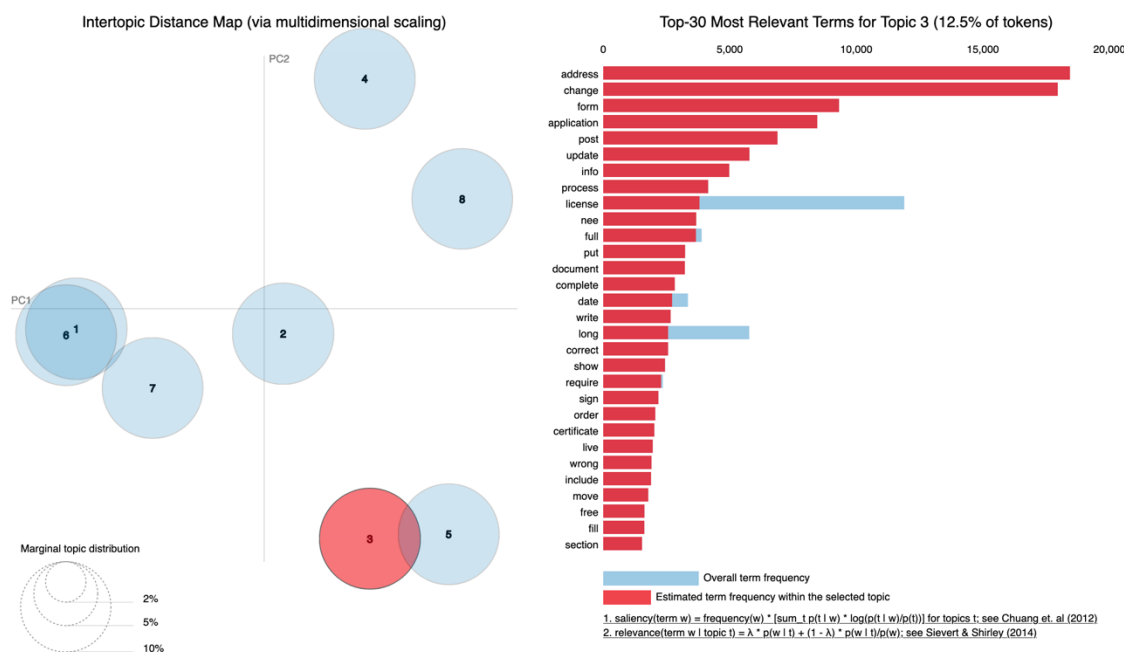


Figure 57 LDA Topic 3

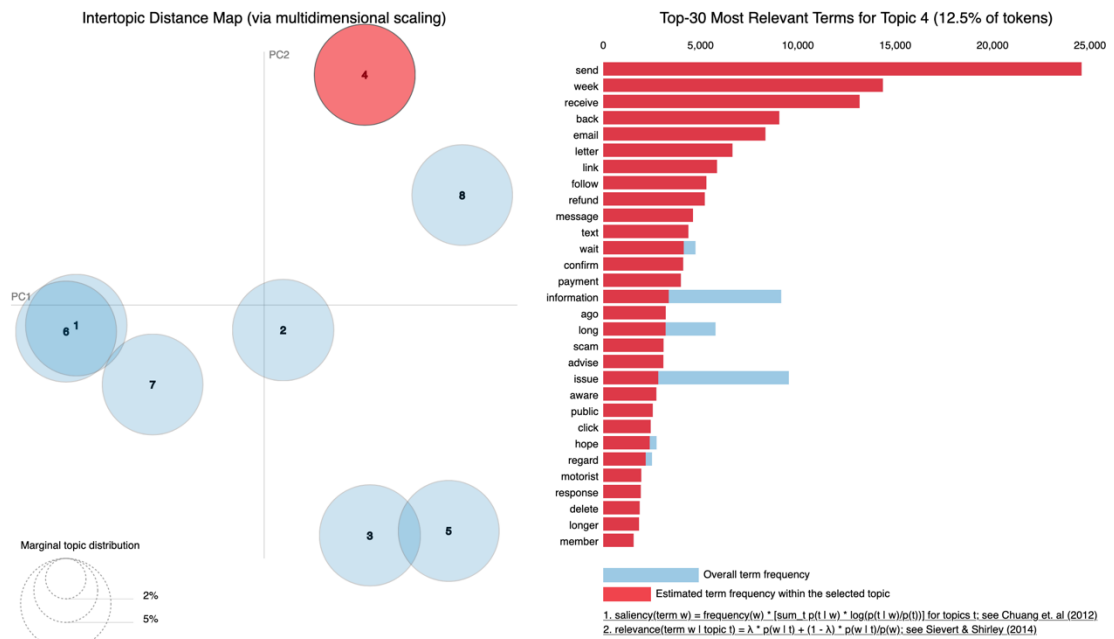


Figure 58 LDA Topic 4

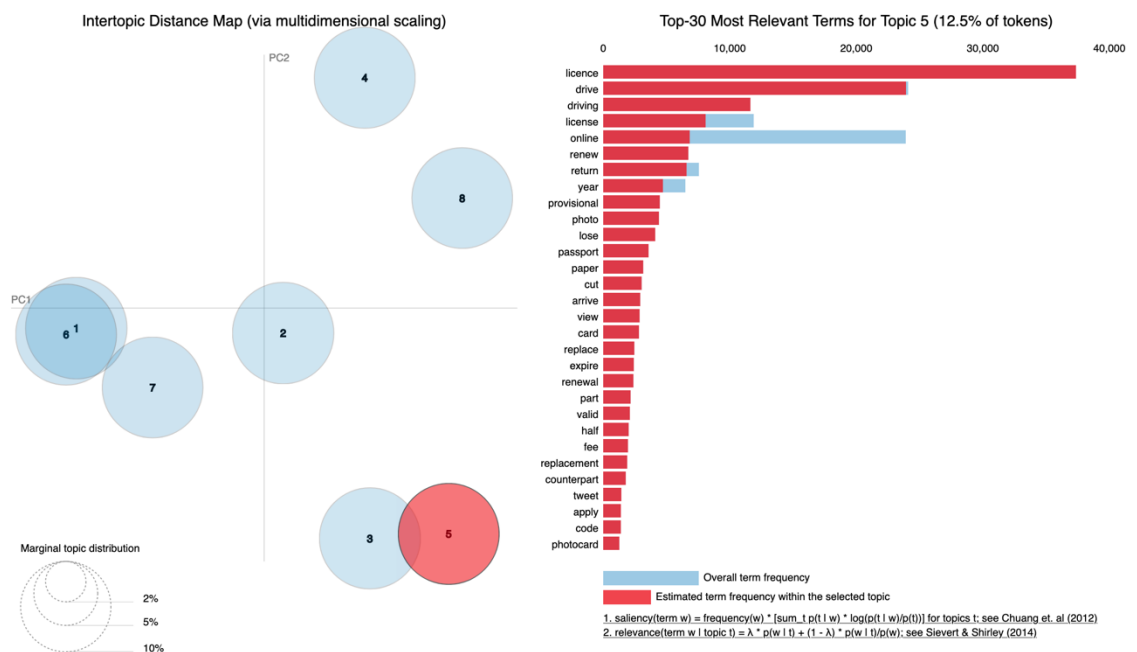


Figure 59 LDA Topic 5

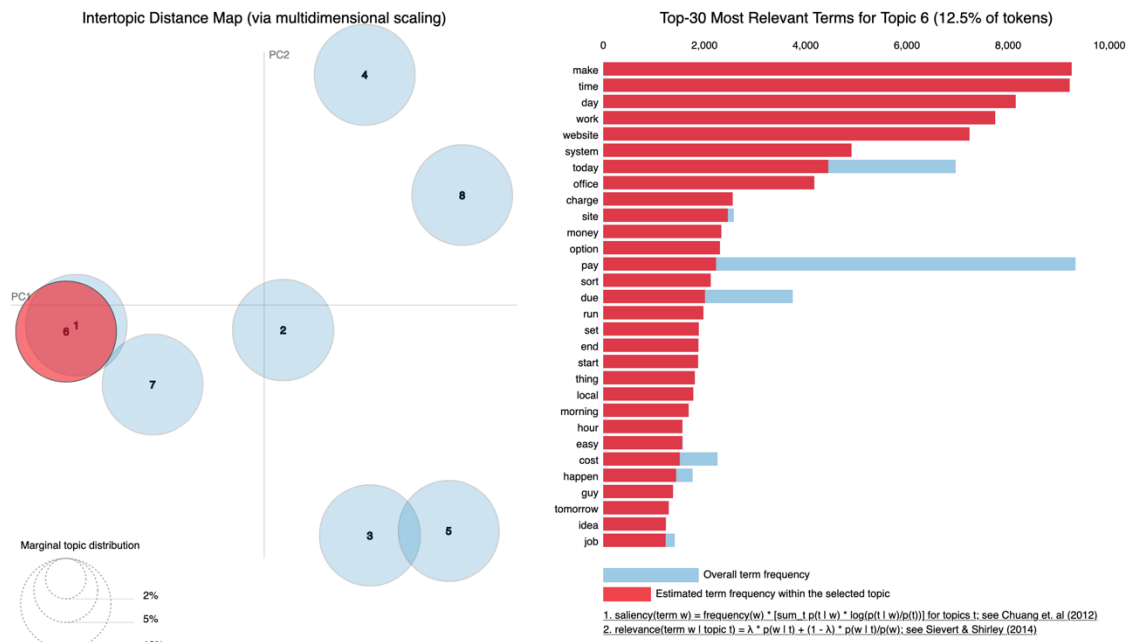


Figure 60 LDA Topic 6

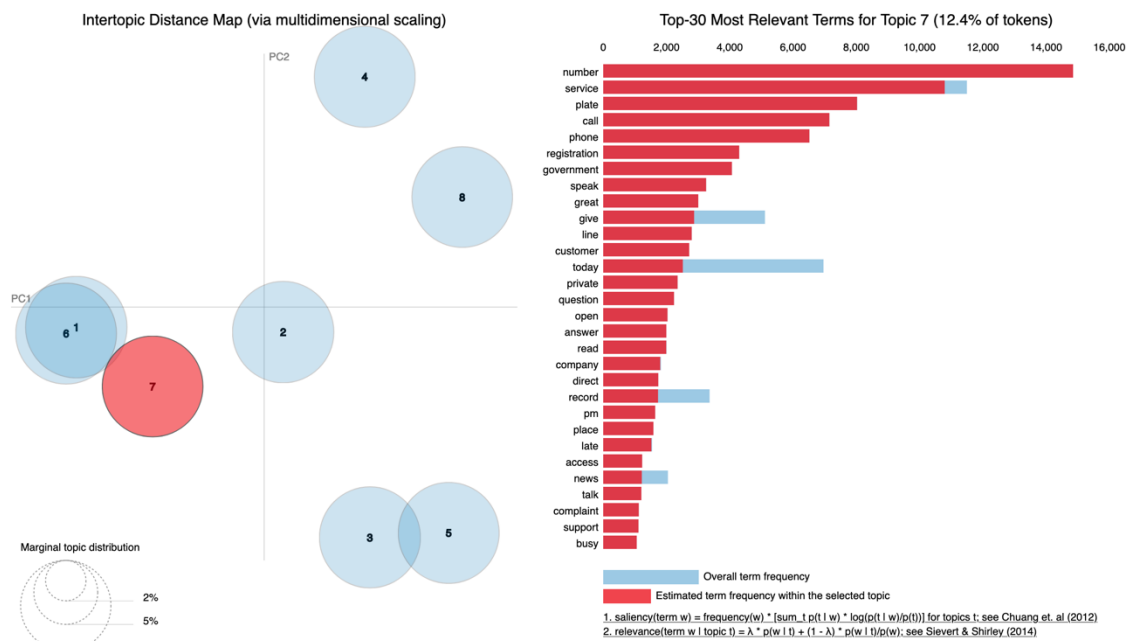


Figure 61 LDA Topic 7

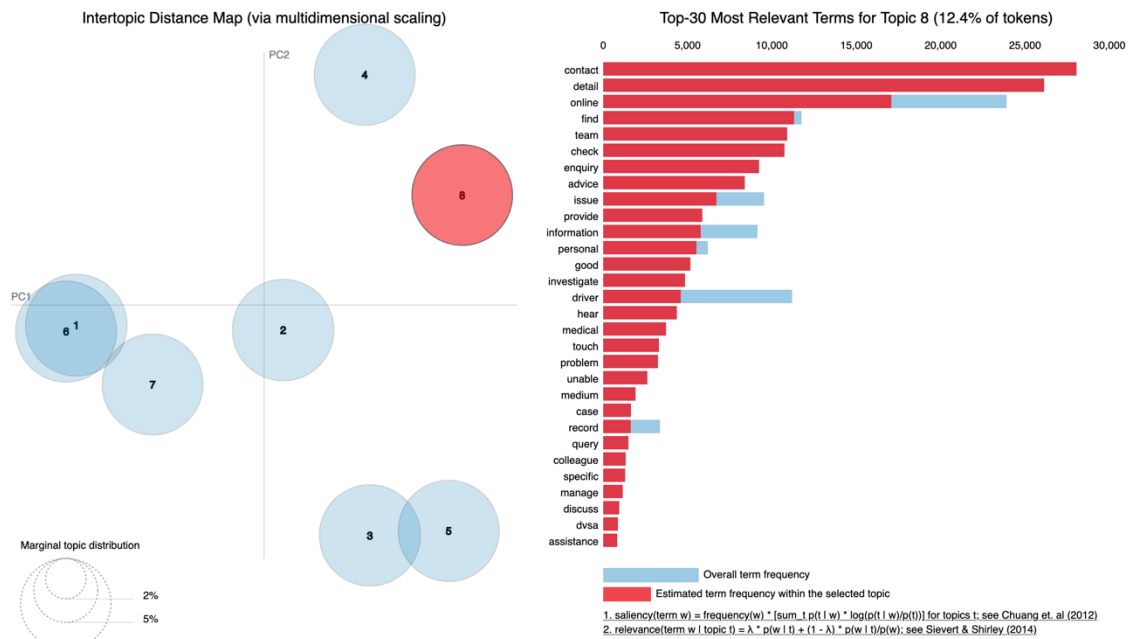


Figure 62 LDA Topic 8

12.3 Development

12.3.1 AWS Lambdas – based in Ohio

12.3.1.1 Sentiment Lambda

```
import json
import nltk
nltk.download('punkt')
from nltk.tokenize import word_tokenize
nltk.data.path.append("/tmp")
nltk.download("punkt", download_dir = "/tmp")
import nltk.tokenize

from vaderSentiment import vaderSentiment as vs
sentimentAnalysis = vs.SentimentIntensityAnalyzer()

def lambda_handler(event, context):
    sentence = event["text"]
    if sentence == "":
        raise Exception("Bad request -- Sentence cannot be null.")
    if len(nltk.sent_tokenize(sentence)) > 1:
        raise Exception("Bad request -- Only one sentence per request.")
```

```

    sentiment = sentimentAnalysis.polarity_scores(sentence)
    return {
      'statusCode': 200,
      'body': sentiment
    }

```

12.3.1.2 Connect to AzureML

```
'use strict';
```

```

const AWS = require('aws-sdk');
const fetch = require('node-fetch');

```

```
exports.handler = (event, context, callback) => {
```

```

  let data = {
    "Inputs": {
      "input1": [
        {
          "Text": event.Inputs.input1[0].Text,
          "Compound": event.Inputs.input1[0].Compound,
          "Positive": event.Inputs.input1[0].Positive,
          "Neutral": event.Inputs.input1[0].Neutral,
          "Negative": event.Inputs.input1[0].Negative
        }
      ]
    },
    "GlobalParameters": {}
  };

```

```
// the bearer key is saved as an environment variable in the settings
```

```

  fetch(process.env.uri, {
    method: "POST",
    headers: {
      "Content-Type": "application/json",
      "Authorization": "Bearer " + process.env.apiKey,
    },
    body: JSON.stringify(data),
  })
  .then((response) =>
    response.json())

```

```

.then((data) => {
  let response = {
    statusCode: 200,
    headers: {
      'Content-Type': 'application/json',
      'Access-Control-Allow-Origin': '*',
      'Access-Control-Allow-Credentials': true,
    },
    body: JSON.stringify(data),
  };

  callback(null, JSON.parse(response.body));

})
.catch((error) => {
  console.error('Error:', error.details);
});
});

```

12.3.1.3 Save results to PostgreSQL

```

exports.handler = async (event) => {
  const { Client } = require('pg'); // Needs the nodePostgres Lambda Layer
  const client = new Client();
  await client.connect();

  const body = event.finalResults;

  const query = {
    text: "insert into public.final
(date,compound,topic1,topic2,topic3,topic4,topic5,topic6,topic7,topic8,assignments,distance0,distance1,distance2) values($1,$2,$3,$4,$5,$6,$7,$8,$9,$10,$11,$12,$13,$14)",
    values:
    [body.date,body.compound,body.topic1,body.topic2,body.topic3,body.topic4,body.topic5,body.topic6,body.topic7,body.topic8,body.assignments,body.distances0,body.distances1,body.distances2],
  };

  const res = await client.query(query);
  console.log(res);
  await client.end();

```

```

const response = {
  statusCode: 200,
  result: "Saved to DB"
};
return response;
};

```

12.4 Results

12.4.1 Sample Text for testing

ID TEXT

- 1 Vehicle tax Web site don't worry it hardly gets used anyway.....
I have been reporting a vehicle for the last 4 month that has no road tax and it is still parked on a main road. The vehicle is classed as SORNED on the DVLA
- 2 website. Will it EVER get clamped?
Is there a reason why you will not just come and clamp this vehicle as it is parked on a main road and it is not taxed, why do you need to correspond with the owner? I am now reporting the vehicle every working day so the report is
- 3 not ignored. I hope you can resolve this issue.
How long does it take to find untaxed cars ? I know one that's had no tax n mot
- 4 since April last year n still running on the road....
there is a car near me that has been used everyday but their mot expired in may
- 5 2018 still being driven today plz inbox for details
So why am I 7 months and still waiting for the dvla to process my application after I have passed a 12 month revoke period, the dvla have been informed no less than 5 times I have no medical issues and this being 19 months after what was basically an ear infection !
The only thing you do fast is remove peoples livelihoods, and the only thing you do cheap is pay a GP £40 to copy a patient's notes and not give a full and honest
- 6 examination of someone's health.
Did this back in August for my son, had to do a medical declaration, but still waiting, can you say how long they take? You've had all info from Drs at the
- 7 beginning of November. Thanks
Done this, he's in the queue. Although before Christmas they told us his file was
- 8 being accessed so it won't be long. Today he's back in the queue.....
I suggest that there processes are out of date or they do not have enough
- 9 staff....they do not care when you call, someone reading a screen.
Ten days since i applied by post for my car tax and it still appears as untaxed on
- 10 your system. Why can't you arrange disabled persons car tax online. Surely

- uploading the necessary paperwork to you is quicker and safer than by post. Or perhaps being able to take into the post office and doing it there. 🙄
- I have already emailed you via your contact page. I am very concerned that my car which I need daily is not taxed and that important personal information may be lost in the post. You do not advise to allow 10 working days on your website for disabled car tax to my knowledge. I await your response to my email which I
- 11 sent yesterday and am advised may take 3 days to answer!
What a leaching government, boasting “we’ll find your car if it’s not taxed”...money, money, we want your money!
Road tax is such an unfair, unbalanced system. If the roads were of decent
- 12 standard it may not be such a slap in the face! 🤡
Whichever way it’s termed, it’s a disproportionate, unfair tax which is demanded in order to drive on the roads and therefore it should be expected that the roads are fit to be driven on.
For example - I had a Discovery 3 which I did high miles in towing, I therefore I also paid a lot of fuel duty...plus the “excise / road tax” was over £500 a year! But say if I’d only used the Discovery one weekend every month I’d still have to pay £500+ to have it parked in my garage!
- 13 That’s how ridiculously unfair this tax is.
its called road tax anymore... It's called vehicle excise duty... Condition of the
- 14 roads and whether you pay your tax because of that is irrelevant...
Yeah, if you have a few cars why could each one be taxed so you can use it on the road. Surely, you can only drive one car at a time and hence the tax could
- 15 be on the driver! Robbers? ???
I never taxed my car For the last 5 year nothing has happened
Lload of rubbish I've reported two such cars in my road no tax, no M.O.T for 4
- 17 years one, the other since October.
Dvla are a pure joke anyways we shouldn’t have to pay road tax it should be like Germany where it comes under ya insurance another way of making money for
- 18 the wealthy
This goverment agency is well past it's sell by date
DVLA clamp it even with tax mot and insurance, in fact if your don’t have all those requirements most people get away with it, they like to clamp the ones
- 20 who are not going to fight them.
A runaround at the local police station not interested and dvla phone call centre theres just no easy and simple way to report finding the untaxed wheel clamps found by a bin area this morning. Numbered BH736 & 740. Cant even add a
- 21 photo here either.
What a laugh, you keep putting up all these ads saying the same, but you've been told of a car for months with no tax n mot n it's still running on the road
- 22 now 9mths later....

Hi there. I have just applied for a driving license ID card (as I still have the paper one :-)) and received a questionnaire to fill in and a list of documents to send in. It doesn't seem clear to what paperwork you need and it actually says that I need to send in my valid passport??? I do not want to send such a document in the post. Is this for real??

Trying to do this for a car I just purchased. just keeps taking me around the same loop and can't progress. tried contacting you by phone but your operators are too busy and no option to hold. please advise

i taxed my car in august for a year . this week you sent me a letter saying its not taxed . Whats going on ? and you have lost my licence again . DVLA useless . bring back the tax disc so we have proof of tax

Been trying to renew my license for a month online but your system doesn't recognise new passports. I found this out after having to phone up and sit through monologue after monologue of information I had already read online, telling me to go online. When I finally got to speak to someone they told me there was an issue and agreed to send me a paper form.

Firstly, there should be a message up on the site stating that some people may have issues with newer passports, and secondly I am furious to be charged a higher renewal fee than if I was to renew online. Plus I have to make a special journey to the post office and buy a postal order plus the postage.

I would like the information needed to make an official complaint about this, as I don't think people should be penalised financially because your IT department are incapable of writing a programme that recognises a combination of 9 numbers. This is an issue you have known about for a while (well over a month) so why has this not yet been sorted?

would you like it if he smacked into your car without tax mot or insurance ? I'm sure you wouldn't, if everyone else has to have their car legal on the road then he shouldn't be any different. Call me what you like if it makes you feel better, but if the shoe was on the other foot, you'd feel the same.....

because they saw fit to make redundancies throughout the UK, blindly believing they could control licensing automatically from Swansea. Previously, they had regional officers who would investigate reports like you mention. All the civil servants involved have since retired on massive pensions or moved to other Gov agencies.

How long does it take for the application of a lost driving license? My mum applied as she's lost her paper one and has never had a photo ID license...she done this on 20th November and still hasn't had anything in the post?!

Sent my licence off to renew 9 weeks ago (plenty of time to get back for ID over Christmas). Chased it up for the last 4weeks to be told TODAY it's been sent to THE WRONG ADDRESS! How Careless can you be with someone's personal documents!

I bet the cheque I sent wasn't paid into the wrong bank account though!

Messed me up going out over Christmas and for New Year, and now someone out there has my personal ID document!

Today (9 weeks later) you tell me you still haven't even got a new one issued to myself yet and it will be soon! Totally unacceptable! What have a paid for-

30 Worry and a headache!

12.5 FrontEnd link

<https://antsite.s3.us-east-2.amazonaws.com/index.html>

12.6 AzureMLaaS Link

`https://ussouthcentral.services.azureml.net/workspaces/5e3857dcc7a3451b8894fbf0eab4e7d0/services/9c69fc985b7f40efbd718cf2db8704c5/execute?api-version=2.0&format=swagger`

REQUEST/RESPONSE	TEST	APPS	LAST UPDATED
	Test Test preview	Excel 2013 or later Excel 2010 or earlier workbook	12/31/2019 12:29:01 AM
BATCH EXECUTION	Test preview	Excel 2013 or later workbook	12/31/2019 12:29:01 AM

Python code:

```
import urllib2
import json

data = {
    "Inputs": {
        "input1":
```

```

        [
            {
                'Text': "ABD Welcomes MPs` Defence Of
Drivers` Privacy - #Parking #DVLA",
                'Compound': "0.4767",
                'Positive': "0.369",
                'Neutral': "0.631",
                'Negative': "0",
            }
        ],
    },
    "GlobalParameters": {
    }
}

body = str.encode(json.dumps(data))

url =
'https://ussouthcentral.services.azureml.net/workspaces/5e3857dcc7a345
1b8894fbf0eab4e7d0/services/9c69fc985b7f40efbd718cf2db8704c5/execute?a
pi-version=2.0&format=swagger'
api_key = 'abc123' # Replace this with the API key for the web service
headers = {'Content-Type': 'application/json', 'Authorization': ('Bearer
'+ api_key)}

req = urllib2.Request(url, body, headers)

try:
    response = urllib2.urlopen(req)

    result = response.read()
    print(result)
except urllib2.HTTPError, error:
    print("The request failed with status code: " + str(error.code))

    # Print the headers - they include the request ID and the
timestamp, which are useful for debugging the failure
    print(error.info())
    print(json.loads(error.read())) // This code requires the Nuget
package Microsoft.AspNet.WebApi.Client to be installed.
// Instructions for doing this in Visual Studio:
// Tools -> Nuget Package Manager -> Package Manager Console
// Install-Package Microsoft.AspNet.WebApi.Client

using System;
using System.Collections.Generic;
using System.IO;
using System.Net.Http;
using System.Net.Http.Formatting;
using System.Net.Http.Headers;
using System.Text;
using System.Threading.Tasks;

namespace CallRequestResponseService

```

```

{
    class Program
    {
        static void Main(string[] args)
        {
            InvokeRequestResponseService().Wait();
        }

        static async Task InvokeRequestResponseService()
        {
            using (var client = new HttpClient())
            {
                var scoreRequest = new
                {
                    Inputs = new Dictionary<string,
List<Dictionary<string, string>>> () {
                        {
                            "input1",
                            new List<Dictionary<string, string>>() {new
Dictionary<string, string>(){
                                {
                                    "Text", "ABD Welcomes
MPs` Defence Of Drivers` Privacy - #Parking #DVLA"
                                },
                                {
                                    "Compound", "0.4767"
                                },
                                {
                                    "Positive", "0.369"
                                },
                                {
                                    "Neutral", "0.631"
                                },
                                {
                                    "Negative", "0"
                                },
                            },
                        },
                    },
                },
                GlobalParameters = new Dictionary<string,
string>() {
                },
            };

            const string apiKey = "abc123"; // Replace this with
the API key for the web service
            client.DefaultRequestHeaders.Authorization = new
AuthenticationHeaderValue( "Bearer", apiKey);
            client.BaseAddress = new
Uri("https://ussouthcentral.services.azureml.net/workspaces/5e3857dcc7
a3451b8894fbf0eab4e7d0/services/9c69fc985b7f40efbd718cf2db8704c5/execu
te?api-version=2.0&format=swagger");

```

```

        // WARNING: The 'await' statement below can result in
a deadlock
        // if you are calling this code from the UI thread of
an ASP.Net application.
        // One way to address this would be to call
ConfigureAwait(false)
        // so that the execution does not attempt to resume on
the original context.
        // For instance, replace code such as:
        //     result = await DoSomeTask()
        // with the following:
        //     result = await
DoSomeTask().ConfigureAwait(false)

        HttpResponseMessage response = await
client.PostAsJsonAsync("", scoreRequest);

        if (response.IsSuccessStatusCode)
        {
            string result = await
response.Content.ReadAsStringAsync();
            Console.WriteLine("Result: {0}", result);
        }
        else
        {
            Console.WriteLine(string.Format("The request
failed with status code: {0}", response.StatusCode));

            // Print the headers - they include the request ID
and the timestamp,
            // which are useful for debugging the failure
            Console.WriteLine(response.Headers.ToString());

            string responseContent = await
response.Content.ReadAsStringAsync();
            Console.WriteLine(responseContent);
        }
    }
}
}
}

```

12.6.1 Swagger API info

