National College of Ireland

Fined-Grained Sentiment Analysis of Yelp Reviews Using Deep Learning Models

MSc Research Project
Data Analytics

Aidan Browne
Student ID: 16140818

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Aidan Browne

Student ID: 16140818

Programme: MSc Data Analytics **Year:** 2020

Module: Research Project

Supervisor: Dr Catherine Mulwa

Submission Due Date: 28/09/2020

Project Title: Fined-Grained Sentiment Analysis of Yelp Reviews Using Deep Learning Models

Word Count: 10,409 **Page Count:** 27

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: 

Date: 28/09/2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□

You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Fine-grained Sentiment Analysis of Yelp Reviews using Deep Learning Models

Aidan Browne
16140818

Abstract

With the development of Web 2.0 capabilities in the early 1990's the way we as human beings interact with each other changed forever. From that period user generated content via social media platforms saw an exponential increase in popularity. Websites such as Yelp became the new place where people discussed how they felt about a business's product or service. Due to this shift businesses more than ever need to understand how the public feel about their product. As part of this project 6 deep learning models were used to make a fine-grained sentiment analysis on the Yelp Open Dataset. In addition to applying sentiment analysis this project attempted to answer if it was possible to increase sentiment accuracy score by selecting reviews considered useful by the public. The models deployed were based on a new technique for Natural Language Processing developed by Google in 2018. With an overall accuracy score an algorithm based on Google's A Light BERT model achieved the best result of 68.39%

1 Introduction

The Internet and its increased accessibility have been described as one of the major characterizing phenomena of present times (Clement, 2020b). With this increased accessibility there has been an exponential change in the way we as human beings interact with each other. As a result, social media platforms have seen a huge growth in popularity as they have enabled users to connect with each more than ever. This increased communication has seen users share their experiences, thoughts and opinions on a vast array of topics from e.g. their feelings about social issues to how they feel about a product (Pathak et al., 2020). Gone are the days when a consumer would find out if a product was worth purchasing through word of mouth. In the age of connectivity, feedback via social media platforms help consumers decide if they should spend their well-earned money. This influence social media platforms have on our spending habits has been described by psychologist Robert Cialdini as The Social Proof Theory. In his book Influence: The Psychology of Persuasion he introduces 6 powers of influence relating to how people perceive themselves or others around them. Social Proof he describes as "we look to our peers for deciding what's acceptable and desirable" (Price, 2011). As a result, individuals are more likely to spend their money on a product if it has received a good review on a site such as Yelp. Subsequently businesses more than ever need to be able to decipher on mass the sentiment the public have towards their products through the analysis of online reviews.

As part of this research project text analytics and deep learning algorithms were applied to the Yelp Open Dataset to predict the star rating of a review of a business.

1.1 Motivation and Background

Online reviews have been described as "voluntary consumer-generated evaluations of businesses, products or services by internet-users who purchased, used, or had experience with

the particular product or service" (Clement, 2019a). In this digital era, they represent the new format of customer feedback which is published online via review websites. They not only represent a written evaluation but sometimes a multi-point scale of 1 – 5 stars associated with user contentment. Sentiment analysis aims to classify this opinion e.g. a movie review as either positive or negative (Pang, Lee and Vaithyanathan, 2002) and this type of sentiment classification is defined as document-level sentiment classification (Liu, 2012). The aim at this level is to classify the overall sentiment of the document through computational linguistics, Natural Language Processing (NLP) and text analytics.

Research carried out in 2019 by Statista showed that in the United States customer reviews accounted for 45% of the public response to how they searched for information on a product they wanted to buy (Kunst, 2020a). This was second only to search engines such as google at 66% (Kunst, 2020a). Further research by Statista also carried out in 2019 indicated that 62% of respondents declared that they found online customer reviews helpful (Kunst, 2020b).

With the evolution of the internet people are spending more time than ever on their mobile phones. A consequence of this has been a massive shift in how people search for information. As of December 2019, 53.3% of web traffic was performed via a mobile device (Kemp, 2020). In the United States alone 69% of adult internet users indicated that they would rather search for product reviews than speak to a shop assistant (Clement, 2019b).

Due to this shift over the last decade the popularity of geo-location review apps such as Yelp and TripAdvisor have grown in popularity. Yelp alone posted net revenue for Q4 2019 of $268.82 million (Clement, 2020c) with over 100 million unique views via mobile web and app for the same period (Clement, 2020a). Independent restaurants that saw a one star increase in their Yelp rating saw an increase in revenue of between 5 – 9 percent (Luca, 2012). With business' nowadays investing a large amount of money and resources into marketing and strategy plans it is essential for them to be able to predict customer attitudes towards their brand or product on mass.

Historically sentiment analysis was generally defined as a binary classification problem i.e. positive or negative. However with the rating-inference problem this led to a more fine-grained approach (Pang and Lee, 2005). The fine-grained prediction problem is complicated as the line between star ratings can sometimes be difficult for even humans to distinguish. In recent years with the development of transfer learning models such as Google's Bidirectional Encoder Representations from Transformers (BERT) by Devlin et al. (2019) and Google Brain and Carnegie Mellon University XLNet by Yang et al. (2019) led to an advancement in star rating prediction accuracy. With this progress one major issue that day to day users or independent businesses still face is the need for computational resources to run these algorithms. As part of this paper, research was carried out to see if users with restricted resources could achieve comparable results by using e.g. the newly released smaller BERT models by Turc et al. (2019) specifically designed for computational restrictions.

1.2 Research Question

The research question was solved by investigating gaps in literature pertaining to the sentiment analysis of online reviews. The outcome of the research was fine-grained Yelp models that can be utilised by users with computational restrictions such as independent businesses. The research question can be defined as follows

RQ:" To what extent can fine-grained sentiment analysis of Yelp reviews be achieved by utilising text analytics and deep learning models (ALBERT, ELECTRA, Smaller BERT,) to predict sentiment star rating when restricted by computational resources".

Sub RQ: "Can the metadata of the review text or user (e.g. review count or user social network) be utilised to increase the quality of reviews and therefore improve prediction accuracy".

1.3 Research Objectives & Contribution

To address the research questions outlined above the below research objectives were specified to achieve the goals of the analysis.

Table 1: Research Objectives

Objectives	Description	Evaluation Metrics
Objective A	Critical review of the existing literature on document level sentiment analysis (2002–2020).	
Objective B	Creation of two separate datasets. One dataset was created with review text and star rating selected randomly. The second dataset was created by utilizing the metadata of the review and user json files to choose review text and corresponding star rating that were considered useful therefore increasing prediction accuracy.	
Objective C	Implementation, results and evaluation of the Fine-Grained Yelp Models which were deployed for each of the two datasets.	
Objective CI	Implementation, results and evaluation of fine-grained sentiment analysis ALBERT model using Yelp dataset 1	Accuracy, Error Rate, MCC
Objective CII	Implementation, results and evaluation of fine-grained sentiment analysis ALBERT model using Yelp dataset 2	
Objective CIII	Implementation, results and evaluation of fine-grained sentiment analysis ELECTRA model using Yelp dataset 1	
Objective CIV	Implementation, results and evaluation of fine-grained sentiment analysis ELECTRA model using Yelp dataset 2	
Objective CV	Implementation, results and evaluation of fine-grained sentiment analysis Smaller BERT model using Yelp dataset 1	
Objective CVI	Implementation, results and evaluation of fine-grained sentiment analysis Smaller BERT model using Yelp dataset 2	
Objective D	Comparison of the six developed Models (Objective C)	
Objective E	Comparison of the six developed Models verses state of the art models	

The research undertaken was a fully developed document level sentiment analysis model based on deep learning techniques. The major contribution that the research achieved was six fine-grained deep learning models outline by Objective C that can be deployed by business owners or everyday users with computational restrictions. This is extremely important for independent businesses as previous research carried out by Luca (2012) showed that an increase in one star rating on the Yelp platform saw a knock on effect of an increase in revenue of 5-9 percent. The minor contribution the research showed was that by selecting useful reviews in turn led to an increase of accuracy score compared to randomly selected reviews. This was true across all three models run during the research.

The remaining sections of the technical report are structured as follows. Section 2 is a review of published literature on document level sentiment analysis between 2002 – 2020. Section 3 proposes a modified CRISP-DM design methodology the design specification process flow and the creation of the two datasets. Section 4 discusses the implementation, results and evaluation

of the three models for dataset 1 & 2. Finally, Section 5 discusses the final thoughts towards the research undertaken and future work recommendation.

2 Literature Review of Document Level Sentiment Analysis of Online Reviews (2002 – 2020)

2.1 Introduction

Over the past number of years sentiment analysis has been defined in different ways by several researchers. The definition as part of this research was introduced by (Liu, 2012) "sentiment analysis, also called opinion mining, is the field of study that analyses people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes".

Research in the field of sentiment analysis relating to online reviews coincided with the advancement of read/write capabilities relating to Web 2.0 in the late 1990's with the growth of blogs, social networks and review sites (Liu, 2012). The aim of the analysis is the classification of an opinion as either positive or negative or a more fine-grained star rating prediction. The literature pertaining to sentiment analysis and the critique of models utilised for fine-grained star prediction will be discussed in the following sections.

2.2 Levels of Sentiment Analysis

From its inception sentiment analysis has grown to be an active research area of NLP. To address the task sentiment analysis is carried out at three levels of granularity (Liu, 2012)

Document Level: as mentioned previously sentiment classification at this level is concerned with the categorization of the sentiment of the whole document as either positive or negative. Analysis at this level assumes that the sentiment expressed by the document is related to a single entity e.g. Yelp review. As a result, a document that expresses sentiment about multiple products cannot be utilised at this granularity (Liu, 2012). Before research carried out by Pang Lee and Vaithyanathan (2002) most of the analysis at this level was concerned with the grouping of documents by their topic e.g. business or sports.

Sentence Level: the sentiment task at this level is to identify the sentiment of a single sentence as either positive, negative or neutral. Subjectivity classification is closely related to sentence level sentiment analysis with the analysis attempting to differentiate between objective sentence and subjective sentences. Objective sentences are sentences that contain factual information while the latter expresses subjective opinions and views. It should be noted that subjectivity does not equate to sentiment as opinions can be implied by objective sentences (Bongirwar, 2015).

Entity or Aspect Level: sentiment analysis at this level is considered more fine-grained analysis than both document level and sentence level. At this granularity the analysis is concerned with the sentiment towards the features or attributes of the product or business that the consumer considers positive or negative. Early research referred to this level as feature-based sentiment analysis and looked at electronic products and the sentiment relating to their attributes such as picture quality and size from Amazon and CNET (Hu and Liu, 2004). With difficult challenges facing both document level and sentence level sentiment classification, aspect sentiment analysis is even more laborious since the analysis entails various sub tasks.

Of these tasks, extraction of the aspects and the classification of these aspects' sentiment are considered two of the most complex undertakings (Liu, 2012).

As the research carried out by the candidate was concerned with the prediction of the star rating of Yelp reviews the rest of the literature review will deal exclusively with research at document level.

2.3 Approaches to Document Level Sentiment Analysis

The main objective of document level sentiment analysis is the classification of an opinion as either positive or negative. From the existing literature it can in be concluded that there are generally two main approaches that have been applied to this problem machine learning and semantic orientation or lexical.

2.3.1 Machine Learning

Machine learning approaches include both supervised, unsupervised or semi-supervised but in general they have attempted to solve two main problems binary classification or regression. A binary classification problem is where the researcher is trying to classify the sentiment as either positive or negative (Chaovalit and Thou, 2005). A regression problem is where the researcher is attempting to predict the star rating associated with the review text (Pang and Lee, 2005).

For supervised or semi-supervised methods, a labelled dataset split into training and test is required to perform the sentiment analysis. After obtaining a good quality dataset the next task is to apply NLP techniques such as text transformation and feature extraction. Text transformation is a very flexible technique and includes converting text to lower case or converting words to their stem e.g. laughing to laugh. Feature extraction refers to the properties of the textual data which are utilised to classify the text such as n-gram tokenization. In computational linguistics n-gram tokenization is where the original text is split into continuous words of e.g. unigram, bi-gram or trigram value. It may also include count vectorization and term frequency-inverse document frequency (TF-IDF) where the occurrences of words are counted, and their frequency given a corresponding weight (Vijayan, Bindu and Parameswaran, 2017). It should be noted that for newer semi-supervised neural network architecture such as transfer learning feature selection is not necessary as many models only require tokenisation (Zhuang et al., 2019). Thereafter a model is trained using an algorithm such as Support Vector Machine (SVM) or Logistic Regression (LR) on the training data and finally evaluated on the test data.

2.3.2 Semantic and Lexicon

Semantic and Lexicon methods relate to the fact that this approach does not require training to predict polarity. Instead the method attempts to assess how much each word is inclined to being positive or negative (Chaovalit and Thou, 2005). Two common approaches are semantic orientation and lexicon.

Semantic orientation attempts to distinguish the direction of each word in a sentence as either positive or negative. Research by Turney (2002) looked at how a sentence has a positive orientation when its associations are good e.g. "great food" or negative orientation when its associations are bad e.g. "terrible movie". The first step for semantic orientation is to apply part-of-speech tagger to the review text. This process takes two successive words from the review text and tries to identify any patterns relating to adjectives, nouns, adverbs and verbs. Next the Pointwise Mutual Information algorithm is used to calculate the strength of the

semantic orientation between the two words. The final step is to calculate an average of the semantic orientation of the sentence with this average number determining if a review is positive or negative and how positive or negative it is (Turney, 2002).

A lexicon-based approach to unsupervised learning sentiment analysis is similarly to semantic orientation but instead uses a dictionary i.e. lexicon of words. This dictionary has been annotated with polarity and strength of each word or words to determine if a sentence is positive or negative. The approach includes intensification and negation handling (Taboada et al., 2011).

2.3.3 Advantages and Disadvantages

Both machine learning and semantic/lexical approaches have advantages and disadvantages. In general, machine learning techniques achieve better results however the trade-off is that they are trained to a specific text corpus and need to be retrained if applied elsewhere. Machine learning techniques can also be prone to overfitting and immensely dependent on the quality of the corpus it is trained on. Although semantic or lexical methods do not achieve as accurate results, they outperform supervised methods in terms to efficiency and latency. As a result these methods are ideal for the implementation of automatic sentiment classification (Chaovalit and Thou, 2005).

2.4 Comparison of Fine-Grained Sentiment Analysis of Online Reviews

2.4.1 Traditional Machine Learning Algorithms

SVM models can be applied to both classification and regression problems and are implemented by constructing a hyperplane or a group of hyperplanes in a high or infinite-dimensional plane to predict the outcome of the task. Prior to the research by Pang and Lee (2005) most of the study in the field of document level sentiment analysis was performed by classifying reviews as either positive or negative. Pang and Lee (2005) believed it was beneficial to possess more information than is provided by binary distinction. The authors concluded that better accuracy was achieved when combining an SVM model with metric labelling. However, the paper showed how complex the task was as the accuracy score achieved for 4 classes was less than achieved for 3 classes.

A RF algorithm as the name implies is a machine learning algorithm based on many decision trees that act as an ensemble model. Each tree in the ensemble calculates an outcome and the calculation with the most occurrences becomes the model prediction. Unlike Pang and Lee (2005), Zhu, Moh and Moh (2016) proposed a multi-layer architecture including a voting classifier. The layers of the architecture consisted of 5 stages including data pre-processing and feature extraction. Rather than defining the issue as a complete 5-star inference problem the authors grouped the task into two levels i.e. datasets. Level 1 included three classes very positive (5-stars), neutral (2,3,4 stars) and very negative (1-star). Level 2 the authors only considered 2,3 & 4 stars as these stars are easily misclassified. After training the classifiers on both levels separately with a voting scheme applied the authors were able to improve the accuracy score compared to Pang and Lee (2005). However, similarly to previous research, improvement was only achieved when the authors split the stars into different levels. To truly achieve a true representation of the rating-inference problem all 5 stars, need to be considered.

Based on Gradient Descent, Stochastic Gradient Descent overcomes the downsides of that model by incorporating randomness. The algorithm is based on the slope of the gradient and the measurement of the degree of change a variable has in comparison to the change of a different variable. The algorithm can also be utilised as part of deep learning methods such as

transfer learning. Moh et al. (2016) used a similar method utilised by Zhu et al. (2016), of a multi-tier architecture and various feature extraction methods to achieve their prediction improvements. However, their research again did not implement a truly fine-grained approach as the authors created 3 datasets consisting of different combinations off Yelp star ratings.

2.4.2 Deep Learning

Deep learning is one of the most exciting subsets of machine learning and has revolutionised the accuracy of fine-grained star prediction. The models are based on neural networks which mimic the neurons found in the human brain and nervous system. The architecture of the network is a multi-layered network of neurons of learnable weights and biases that can be utilised for classification or regression problems. The multiple layers consist of an input layer, an output layer and then hidden layers in-between made up of the neurons.

Convolutional Neural Networks (CNN's) are a type of neural networks that were first developed for image classification however they are now implemented for NLP tasks. The models are based on a feedforward neural network made up of convolutional and pooling layers. The convolutional layer is the main constructing block of the neural network which combined with an activation function such as Rectified Linear Unit (ReLu) is utilised to model a non-linear relationship. Zhang et al (2015) implemented a character-level convolutional network. The design of the model was modular with the gradients achieved through backpropagation. Unlike previous research the authors implemented a fine-grained sentiment analysis using Yelp. The authors created a dataset consisting of 700,000 reviews, 650,000 for training and 50,000 for testing equally distributed between each star. The dataset was a subset of the Yelp Data Challenge 2015 and has been utilised by preceding research that used Yelp data. One of the issues with the dataset was the authors did not outline how it was to be created for when future releases of Yelp data were released.

Finally Johnson and Zhang (2017) implemented a variation of shallow word level CNN's which included features of down sampling, shortcut connections and text region embedding.

Transfer Learning is based on the idea that we as human beings can apply knowledge learned from one task to accomplish another. However traditional machine learning and other deep learning algorithms were designed to learn in isolation. Therefore, transfer learning was designed to overcome this paradigm of isolated learning and exploit the transfer of knowledge to complete other tasks. To accomplish these tasks different learning strategies and techniques can be applied to the sentiment analysis problem. These include Inductive Transfer Learning, Unsupervised Transfer Learning and Transductive Transfer Learning (Pan and Yang, 2010).

Research carried out by Sun et al. (2019) and Munikar et al. (2019) utilised the BERT model developed at Google by Devlin et al. (2019). The BERT model applied an unsupervised learning strategy and was pre-trained on a large unlabelled text corpus of English Wikipedia and BooksCorpus datasets. Thereafter it can be fine-tuned for downstream NLP tasks such as text classification with a labelled dataset. The model introduced Masked Language Model (MLM) and Next Sentence Prediction (NSP). For fine tuning Sun et al. (2019) implemented various fine-tuning strategies with the optimal model based on in-task pretraining. Similarly to Sun et al. (2019), Munikar et al. (2019) implemented finetuning using the Stanford Treebank fine grained dataset SST-5.

Research carried out by Howard et al. (2018) developed a model based on inductive transfer learning called Universal Language Model Fine-tuning (ULMFiT). The model differed to BERT as the source and target domains contained labelled data. It achieved state-of-the-art (SOTA) results by utilising innovative techniques of discriminative fine-tuning, slanted triangular learning rates and gradual unfreezing.

Finally, research carried out by Yang et al. (2019) similarly to Devlin et al. (2019) applied unsupervised transfer learning techniques but utilised the best of autoregressive language modelling and autoencoding. The model was called XLNet.

One of the major issues concerning the above-mentioned models was the required access to large quantities of computational resources due to their avaricious appetite. To run the models from pre-training took days with BERT-large alone pre-trained on 16 cloud TPU's (64 TPU chips) over 4 days. As a result, research was undertaken to reduce model size and accelerate its running time while upholding model accuracy. Some of the approaches included Parameter Reduction (PR), Replaced Token Detection (RTD) and Knowledge Distillation (KD).

Models that have optimized the architecture of the BERT to reduce the number of parameters include A Lite BERT (ALBERT) by Lan et al. (2019). The researchers reduced training time for ALBERT by introducing Factorized Embedding Parameterization (FEP), Cross Layer Parameter Sharing (CLPS) and Sentence Order Prediction (SOP). Similarly, to BERT, ALBERT was pretrained using the BOOKCORPUS and English Wikipedia datasets.

ELECTRA on the other hand optimized training time by introducing RTD for pre-training and was trained on uncased English text. Both models achieved competitive results against the General Language Understanding Evaluation (GLUE) benchmark but only achieved SOTA results when they were scaled up with ALBERT-xxlarge and ELECTRA-large models. Even though both models were successful in reducing training time the two models still require considerable computational resources to run the largest models due to their size. ALBERT-xxlarge had 233m parameters and ELECTRA-large had 335m which is comparable to BERT-large.

Knowledge Distillation (KD) is a type of model compression technique in which a smaller student model is trained to replicate a larger teacher model. By doing so it attempts to reduce the model size and accelerate its running time while achieving comparable results to the larger teacher model. Models based on KD include smaller BERT models by Turc et al. (2019) and FastBERT by Liu et al. (2020). Turc et al. (2019) followed three accepted procedures of MLM, task-specific distillation and fine-tuning. FastBERT however consisted of architecture of backbone and branches and achieved competitive accuracy scores on the Yelp Dataset compared to BERT-base and DistilBERT.

2.5 Comparison of results of Fine-Grained Sentiment Analysis

A comparison of models utilised for fine-grained sentiment analysis using the Yelp dataset are listed below.

Table 2: Comparison of Fine-Grained Sentiment analysis Models

Dataset	Model	Error Rate	Authors
Yelp 5-Star	CNN	37.95%	(Zhang et al., 2015)
Yelp 5-Star	DistilBERT	35.75%	(Liu et al., 2020)
Yelp 5-Star	BERT-base	34.07%	(Liu et al., 2020)
Yelp 5-Star	FastBERT	34.07%	(Liu et al., 2020)
Yelp 5-Star	DPCNN	30.58%	(Johnson and Zhang, 2017)

Table 2 (continued): Comparison of Fine-Grained Sentiment analysis Models

Dataset	Model	Error Rate	Authors
Yelp 5-Star	ULMFiT	29.98%	(Howard and Ruder, 2018)
Yelp 5-Star	BERT-Large+ITPT	28.62%	(Sun et al., 2019)
Yelp 5-Star	XLNet	27.80%	(Yang et al., 2019)

2.6 Quality of Reviews

The abundant number of reviews available for consumers to evaluate if they want to purchase a product or service does not come without drawbacks. With every high-quality review that conveys useful information there is also a plethora of fake reviews, opinions that don't benefit the consumer and ones that are very misleading and highly subjective (Lu et al., 2010). Due to this review sites such as Yelp introduced ranking votes of useful, funny and cool so that reviewees could vote if a review conveyed worthwhile information. The aim was to tackle fake reviews and ensure reviews of a high quality came to the forefront. However, this did not completely solve the issue. "There is a rich-get-richer effect where the top reviews accumulate more and more ratings, while more recent reviews are rarely read and thus, not rated" (Lu et al., 2010). To tackle this problem Lu et al. (2010) predicted review usefulness utilising a combination of text and social context information to determine the quality of the reviews. This included the total number of tokens, total number of sentences and the number of past reviews by the reviewer. The authors utilised a linear regression model to predict review quality. Kim et al. (2017) utilised a similarly approach to Lu et al. (2010). The authors considered aspects of content (informativeness, sentiment, readability), source (reputation, geographical entropy) and business (star rating, number of reviews, if open or closed) to predict the useful ranking score. A review was classified as useful if it had 5 or more useful votes. The authors utilised LR with features based on intrinsic reputation, content informativeness and rating informativeness.

2.7 Gaps in Research

As demonstrated by research by Yang et al, (2019) improvements in the accuracy of fine-grained sentiment analysis accuracy has been achieved by applying deep learning techniques. However, for everyday users and independent businesses the models developed require access to considerable computation resources. Due to this newer model's based on KD, PR and RTD were developed e.g. ALBERT. These newer models to the best of the candidate's knowledge have only been measured against the GLUE benchmark and therefore, research was undertaken utilizing the Yelp Open dataset. Also, the quality of the review text has an impact on the accuracy gained by a model. Therefore, also to the candidate best knowledge most of the research has been conducted by either predicting star polarity or usefulness score separately. As a result, research was also carried out by the candidate to see if by combining both approaches of selecting useful reviews star prediction accuracy can be increased.

In the following section the methodology used to achieve the research objectives is discussed.

3 Yelp Methodology Approach and Design Specifications

3.1 Fined Grained Sentiment Analysis Yelp Methodology

For the successful implementation of any data mining project it is necessary to follow a structured methodology that is robust and well-proven in approach. There are three generally accepted approaches for data mining projects, Cross-Industry Standard Process for Data Mining (CRISP-DM), Knowledge Database Discovery (KDD) and Sample, Explore, Modify Model, Assess (SEMMA) (Azevedo and Santos, 2008).

As part of the project the candidate utilized a modified CRISP-DM approach for the fine-grained sentiment analysis of Yelp online reviews. The below figure illustrates the modified approach which follows a cyclical nature.

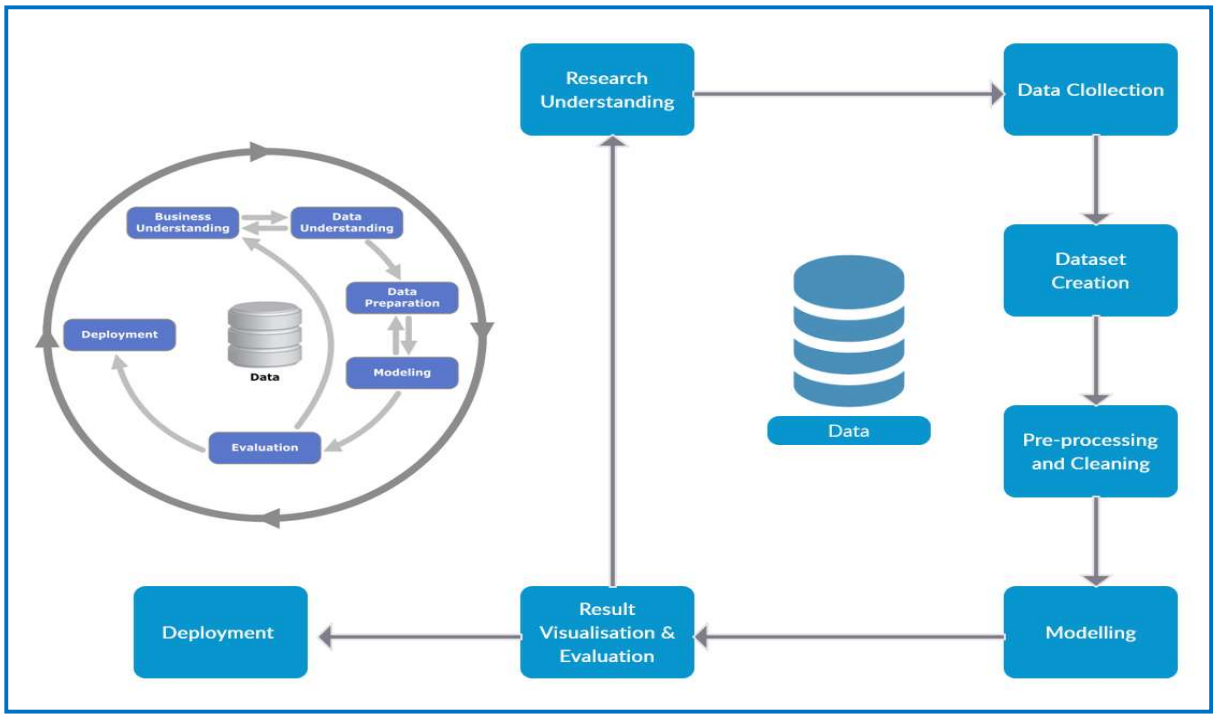

Figure 1: Yelp Modified CRISP-DM Methodology

Established in 1996 the cyclical process follows various stages for the development of a data mining project

- Research Understanding: the first phase of the research involves the understanding of objectives and specifications of the data mining project related to the business perspective i.e. fine-grained sentiment analysis of Yelp reviews
- Data Collection: the next phase deals with the collection of the data which is required to meet the objectives of the research project. During this stage the required json files are downloaded from the Yelp Open Data website[1].

[1] https://www.yelp.com/dataset

- Dataset Creation: at this stage all activities required to create the two datasets were performed. This involves loading the Business, User and Review json files to Azure Databricks for distributed processing using PySpark
- Data Pre-processing and Cleaning. At this stage the datasets are transformed to be able to be utilised by the Fine-Grained Yelp Models. The process occurs at two different stages. Before the datasets are transferred to the candidate's local drive from Databricks new lines, carriage returns, and all unnecessary characters are removed. Next the two datasets are transformed into the format required by the Simple Transformers library in Google Colaboratory (Colab).
- Fine-Grained Sentiment Analysis Implementation: during this step the 6 models chosen for the task of sentiment star prediction were run on each dataset.
- Result Visualisation and Evaluation: while the models were running, they were tracked against recognised scientific measures using the Weights & Biases (W&B) website. W&B is a tool for tracking machine learning experiments in real time with the results achieved visualised using Tableau. At this stage it will be determined if the data mining project has achieved the objectives.
- Deployment: if the project has been determined successful the final stage involves the strategy for deployment of the optimal model.

3.2 Design Process Flow

A multi-tier design architecture was implemented for the classification of the Yelp star rating. The architecture consists of three tiers. The first tier is the client layer which relates to the user interface, experiment tracking and data visualization of the results. The second tier relates to the business layer consisting of the functional business logic tier of the deep learning models ALBERT, ELECTRA and Smaller BERT. Finally, the third tier the data layer, contains the back-end processing cloud platforms and data. Cloud platforms utilized during the project include Azure Databricks, Google Drive, Colab, JupyterLab and Genesis Cloud. The three-tier architecture is represented in below diagram.

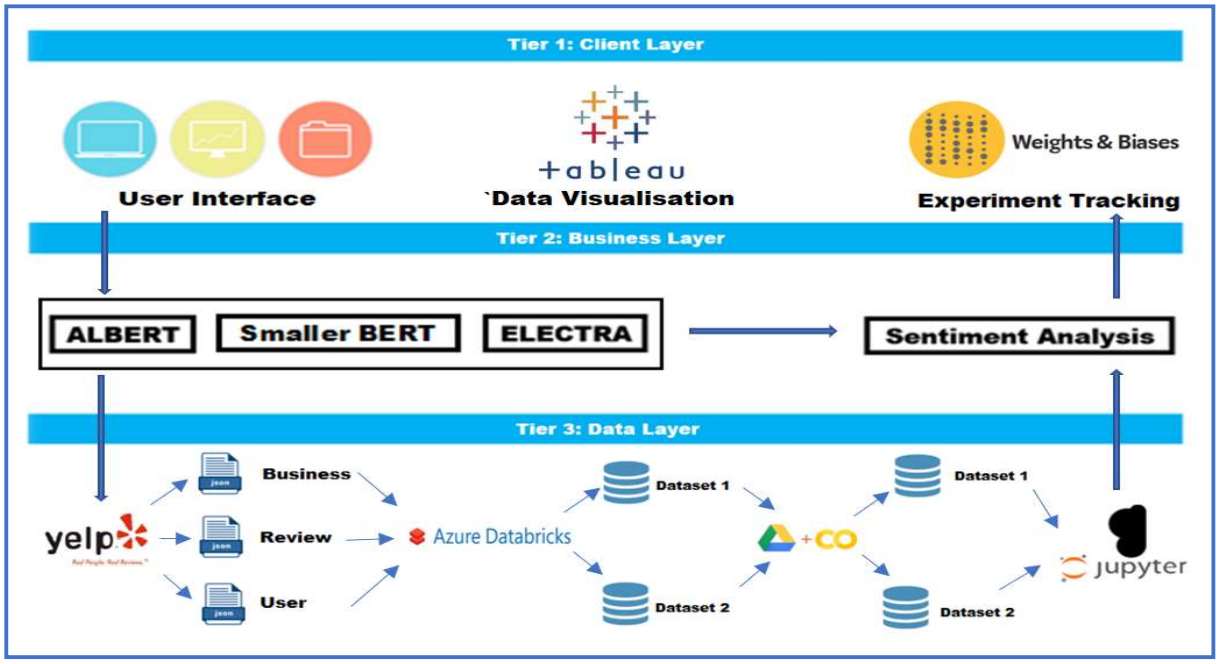

Figure 2: Three-tier Architecture

3.3 Dataset Creation Process

After obtaining the three datasets the JSON files were uploaded to Azure blob containers from the candidate's local drive. Thereafter they were mounted to Databricks for processing with PySpark to create the two datasets. The process involved in the creation of the two datasets is described below.

3.3.1 Dataset 1

Dataset one was created by applying stratified sampling to the review DataFrame with 700,000 reviews selected equally distributed between each star rating. After stratified sampling the review DataFrame was merged with the business DataFrame.

3.3.2 Dataset 2

Dataset two was created by applying a LR model to predict if a review was useful or not. By doing so the candidate attempted to overcome the rich-get-richer effect previously mentioned in Section 2. By selecting quality reviews, the hypothesis was that prediction accuracy of star rating would be higher compared to Dataset 1. The process followed to create Dataset 2 was as follows.

The first step in the process is to create the features to be used by the LR model. The features selected and created fell into 3 categories social network, content informativeness and review ratings.

Social Network refers to the relationship that a user has with other users throughout the Yelp platform. The hypothesis is that a user with a high social ranking is trusted and therefore the reviews they leave are informative and of a higher quality. To calculate the social network of a user the GraphFrames library in PySpark was used. The two algorithms calculated were outDegree and PageRank. The outDegree algorithm calculates the number of edges (friends) that are going out from a vertex (user) in a directed graph. The PageRank algorithm was developed by Google to rank the relevant importance of a web page. PageRank is a vote by all the other user's on how important a user is. A link to a user is counted as declaration of support. The presence of no connection is a vote of no support. Also included in social network category is the number of past reviews a user has written.

Content informativeness relates to the textual statistical features of the review text. The textual statistics that were calculated include word count, sentence count, price count, exclamation count, question count and average sentence length. The hypothesis is that an informative review will be made up of higher word and sentence count. The use of punctuation will also be used in a superior review. Finally, a review that conveys a price description is one that conveys informative information that a reviewee would find helpful.

Review rating contained in the DataFrames were classified under three types, average business rating, average user rating and the rating for each review. Due to the abnormality caused by the deviation between the different ratings the following 3 ratings were included. The average review rating of a review, the absolute difference between the average review rating and the average business rating. Lastly the absolute difference between the average review rating and average user rating. A description of the features is outlined in the below table.

Table 4: Feature Description

Feature Name	Type	Feature Description	Dataset
Social Network			
PageRank	Social Network	PageRank of user	User
Out-Degree	Social Network	Out-Degree of user	User
Review Count	Author	Number of past reviews by user	User
Content Informativeness			
Word Count	Text-Statistic	Number of words in a sentence	Review
Sentence Count	Text-Statistic	Number of sentences in review	Review
Price Count	Text-Statistic	Count of number of times that price is mentioned in a review	Review
Exclamation Count	Text-Statistic	Count of number of times that a question is proposed in a review	Review
Question Count	Text-Statistic	Count of number of times that an exclamation is used in a review	Review
Avg. Sentence Count	Text-Statistic	The avg. number of words in a sentence in a review	Review
Review Rating			
Avg. Review Rating	Review Rating	Avg. rating of a review	Review
Deviation to avg. business rating	Review Rating	Abs return of (avg. review rating) – (avg. business rating)	Review & Business
Deviation to avg. user rating	Review Rating	Abs return of (avg. review rating) – (avg. user rating)	Review & User

After the features were generated, they were merged with the review DataFrame which contained the date and useful vote count of each review for further processing.

Next it has been shown that review exposure time has an influence on the number of useful votes a review receives. Therefore, it was decided to filter the DataFrame for a 4-month time period to mitigate for this and to ensure that recent reviews were included in the training of the model. The time period selected was from 01/11/2018 - 31/01/2019.

Following this a user defined function was applied to assign a label of either 1 or 0 to each review. Reviews that had 5 or more useful votes were converted to 1 and reviews with less than 5 votes converted to 0. Stratified sampling was applied to generate an equally balanced DataFrame containing 21,361 reviews for each label.

Next multicollinearity was checked. Multicollinearity in regression refers to the occurrence of correlation between the independent variables in a model. To test for multicollinearity a correlation matrix utilizing the Seaborn Heatmaps library was produced. Word count and sentence count had a high positive correlation of +0.8. It was decided to drop sentence count. The correlation between the dependent variables and the independent variables was also checked. There was no significant correlation between the dependent variable and the independent variables which ranged between -0.11 and +0.38.

Next data preparation and feature engineering was applied. First the DataFrame is checked for missing values. With no missing values in the DataFrame feature engineering was applied to the DataFrame. The process involves firstly selecting all the independent numeric features and converting them to one single feature vector using VectorAssembler. Next StandardScaler was applied to the feature vector. This normalised the feature vector by subtracting the mean and dividing by the standard deviation of each feature. The DataFrame was split 80% train 20% test

After data preparation and feature engineering, feature importance was performed using a RF classifier run over 20 trees. The results of the feature importance model indicated that 4 features were responsible for 91% of the outcome of the RF model. They were review_count (0.3996), outDegree (0.2187), wordCount (0.1814), and pagerank (0.1066). The LR model was run using the 4 features and achieved an accuracy score of 0.81 and a F1 score of 0.79

With the model now developed the last step in the creation of Dataset 2 involved deploying the LR model to the full review DataFrame to select which review to include in Dataset 2. The above steps of label conversion, vector assembler, standardization were applied to the four features of significance outlined above. After the model was run the DataFrame was filtered for review id's where the predicted outcome equalled 1. The prediction DataFrame was merged with the review and business DataFrames. Stratified sampling was again applied to create a DataFrame of 700,000 reviews balanced equally for each star rating.

3.4 Data Pre-Processing and Exploration

3.4.1 Data Pre-Processing

Data pre-processing occurs at two stages. Firstly, after the two datasets were created the text data was cleaned of white spaces and characters especially commas so that there would be no errors caused when converting to csv. After this was complete the files were transferred to the Databricks File System (DBFS) to be exported to the local drive before they were uploaded to Google Drive.

After the datasets were transferred to Google Drive, they were processed further in Colab so that they were in the correct format for the Simple Transformers Library. The format required for multi-class sentiment analysis is a two-column dataset made up of the text and stars columns. As tokenisation is done by the model the text does not need to be processed further. For the star rating the models require that the review star to start at 0 and therefore, each star was reduced by 1. The datasets were split training (600,000), validation (50,000) and test (50,000). It should be noted that when creating the original datasets in PySpark using stratified sampling an approximation of the fraction is returned when using this method. This approximation can be overcome by using sampleByKeyExact() but is currently unavailable for Python users when using Spark (*MLlib - Basic Statistics*, no date). Therefore, there is a slight difference in the datasets with regards to the distribution of the star ratings. When created Dataset 1 contained 699,985 reviews and Dataset 2 contained 700,004 reviews. As the difference is not significant it was decided to proceed using both datasets. After the two datasets are split into training, validation and test, they are saved again in Google Drive before running the sentiment analysis in Genesis Cloud with JupyterLab.

3.4.2 Data Exploration

One of the limitations of the models used as part of the research was that they can only accept a max of 512 tokens. To save substantial memory it is advisable when fine-tuning to use a shorter token count as pre-training is done using the max token count. To check the token distribution of each dataset a sample of 500,000 reviews were selected for data exploration in Colab. The full dataset was not used as the size of the dataset caused Colab to freeze when running on the candidate's laptop. The BERT Tokenizer was imported from the Transformers library and a function created to be applied to each dataset. The tokenizer function splits the sentences into tokens, adds the classification token [CLS] at the start of the sentence and a special token [SEP] at the end of the sentence and maps ID's to each token. When the function is run an error is reported for every sequence that contains more than 512 tokens. After the

function was run on each dataset the min, max, mean and median token length of each Dataset was calculated. Dataset 1 had a min (3), max (3,090), mean (150), median (109). Dataset 2 had a min (3), max (2,795), mean (234), median (198). Dataset 1 had 2.6% of sentences over 512 tokens while dataset 2 had 6.5%. The token distribution of both datasets can be seen in below figures.

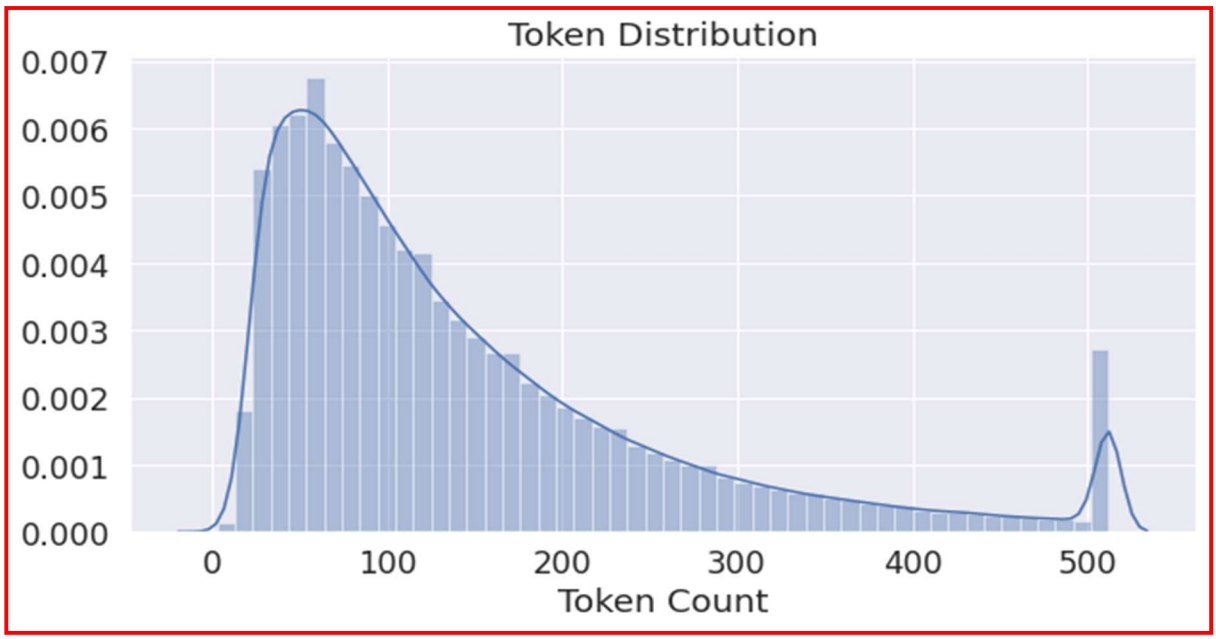

Figure 3: Dataset 1 Token Distribution

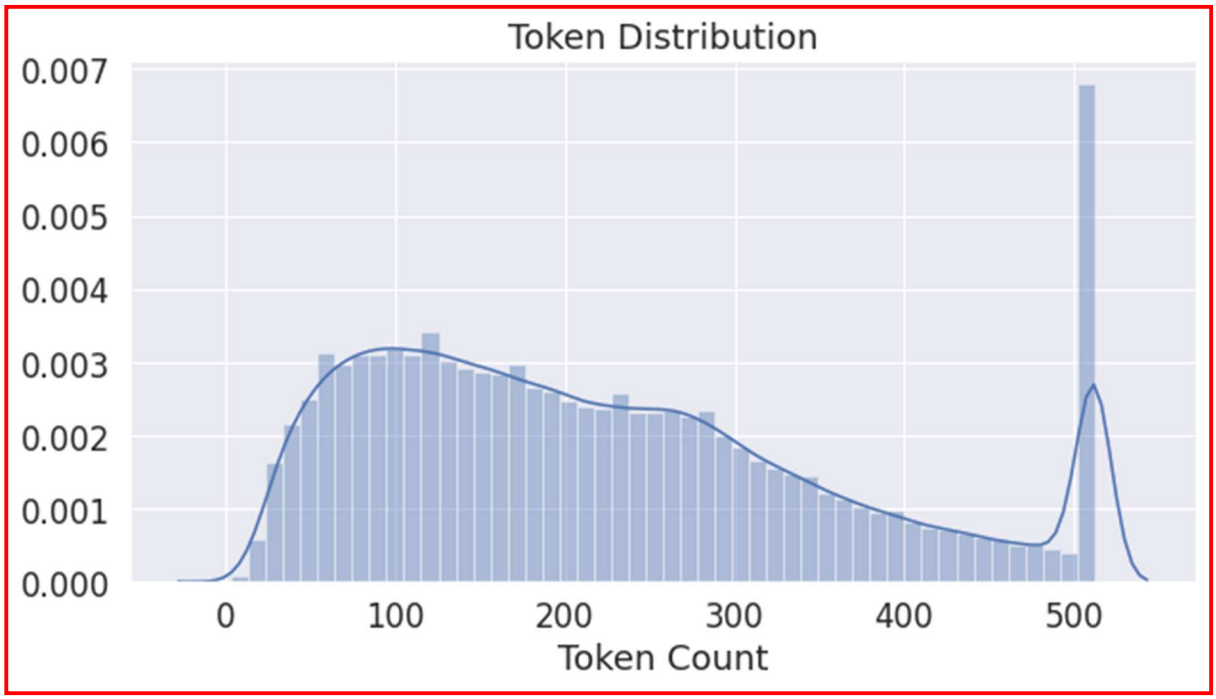

Figure 4: Dataset 2 Token Distribution

At this stage it was decided to run the models for Dataset 1 at a max of 150 tokens and Dataset 2 at max of 235 tokens which equate to the average length for both datasets. This was due to

both datasets having only a small percentage of sentences with tokens over 512 and to save on computational resources.

After Data Exploration was complete the next stage of the research was the implementation of the chosen deep learning models. The following section details model selection and their implementation with the Simple Transformers library, the result metrics chosen and the evaluation of the most appropriate model for the purpose of the fine-grained sentiment analysis.

4 Implementation, Results and Evaluation of the Fine-Grained Yelp Models

4.1 Sentiment Analysis with Simple Transformers

After formatting the two datasets the next step was to apply the Transformer models to achieve the goal of fine-grained sentiment analysis. One of the major benefits of Transformer models are that they are based on transfer learning. As a result, the pre-training process which usually is the most costly and time-consuming part is only required to be performed once and thereafter the same model can be applied again with fine-tuning for different tasks. To date the models developed by Google and other researchers have been made public via GitHub or the Hugging Face library. During the research undertaken by the candidature the Simple Transformers library was used to run the models. The Simple Transformers Library was built on top of the Hugging Face library for the easy implementation of Transformer models.

Even though the Simple Transformers library has simplified the implementation of the Hugging Face library access to Graphic Processing Units (GPU's) is still essential in running the models. To overcome the candidates limited access to computational resources JupyterLab and Genesis Cloud were utilized. Genesis Cloud is a cloud GPU service that offers Nvidia GeForce GTX 1080 Ti at cost effective prices. At the time of undertaking the research Genesis Cloud was in Beta Phase II and were offering GPU's at a 50% reduction in their hourly rate. To accomplish the research two GTX 1080 Ti were used. This is the default allocation of the quota assigned to each user on registration. If a user requires access to more than two GPU's they are required to log a ticket with the Genesis Cloud support team.

4.2 Overfitting

As neural networks are prone to overfitting evaluating during training was essential in preventing the model in producing inaccurate results on the test data. This was achieved by using a validation or development dataset when training combined with early stopping. Early Stopping is a technique where if the model does not improve over several evaluation steps the model stops running. When configuring the Simple Transformers library, the models were set to terminate after five evaluations if there was no increase in the Matthews Correlation Coefficient (MCC).

MCC is applied in machine learning as a classification measure for binary or multi-class classification tasks. MCC as the name applies is a measurement of the correlation coefficient similarly to the calculation of the correlation between two variables. The coefficient ranges from -1 to +1 with +1 indicating perfect correlation. On the other hand, a model closer to -1 indicates that the model is not accurate in predicting the outcome variables. The metric was chosen as it provides a balanced evaluation of the predicted outcomes compared to evaluation derived from percentages such as accuracy (Baldi et al., 2000). The equation for MCC is defined below with True Positive defined as TP, True Negative as TN, False Positive as FP and False Negative as FN.

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Equation 1: Matthews Correlation Coefficient

4.3 Results Metrics

When the model was finished running it was evaluated again using the test dataset with accuracy calculated for the selection of the final optimal model. After accuracy had been calculated for both the validation and the test data it would have become apparent if the model was overfitted as there would have been a significant difference between these two evaluation points. Finally, for comparison to previous research the Error Rate was also calculated for the test results.

Accuracy can be defined as the number of correct predictions divided by the total number of predictions. Error Rate refers to how often the model is incorrect in predicting an outcome. Error Rate is calculated by subtracting Accuracy from 1. The equation for accuracy is defined below

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Equation 2: Accuracy

4.4 Overview of Bidirectional Encoder Representations from Transformers

As previously mentioned in Section 2 research was undertaken to reduce the size of the BERT model as a user with restricted computational resources could not run BERT-large which had 344 million parameters. BERT was developed in 2018 and was a new method of language representation that used bidirectional training of a Transformer. A Transformer is a common attention model which attempts to understand the contextual relationships amongst words in a sentence. The architecture or learning strategy consisted of encoder layers with multi-head attention and bidirectional series input. The encoder layers read the input text and bidirectional referred to the model's ability to see how a word fitted into the context of all the other words in a sentence. The model saw the word before and after a word in a sentence and had two self-supervised objectives of Masked Language Model (MLM) and Next Sentence Prediction (NSP).

In MLM a sequence of words is digested by BERT with 15% of the tokens masked at random. BERT then attempts to predict the masked tokens by understanding the context of the unmasked tokens. Masked tokens are utilised in bidirectional models due to information escaping at the lower layers which allows a token to see itself in the subsequent layers.

Next NSP attempts to predict if given a pair of sentences that the second sentence is the true next sentence for the first sentence. During training the model first selects 50% of the correct sentence pairs with the remaining sentence pairs selected at random. The model adds two tokens to distinguish between the two sentences. It adds a classification token [CLS] at the start of the sentence and a special token [SEP] at the end of the sentence.

The results of the paper showed that a model trained on bidirectionality can show a deeper understanding of natural language compared to models based on single direction.

4.5 Implementation, Evaluation and Results of Albert-base Model

4.5.1 Implementation

ALBERT was developed by Google Research and Toyota Technological Institute paper ALBERT: A Lite BERT for Self-supervised Learning of Language Representations (Lan et al., 2019). Version 1 was released on 26th September 2019 and version 2 which has been used as part of this research was released on the 9th February 2020. As previously mentioned, ALBERT made three considerable but crucial changes in comparison to BERT with the introduction of FEP, CLPS and SOP.

For FEP the researchers noted that the embeddings size had a direct correlation to the size of the hidden layers. The WordPiece Embeddings referred to the vocab size of 30,000 words (same as BERT) and the vectors that represented them. The functionality of BERT was tied to its ability to learn context dependent representations by way of the hidden layers. However, this led to models with millions of parameters. The authors discovered by applying factorization to the embeddings parameters their size could be significantly reduced e.g. ALBERT-Base 12m parameters compared to BERT-Base 108m.

With CLPS the effectiveness of the parameters was further increased with the sharing of parameters over all layers. Lastly ALBERT attempted to predict SOP rather than NSP. In SOP the model takes two sentences from the same document and classifies these as in the correct order. Then it swaps the sentence order and classifies this as incorrect. By doing so it is attempting to learn a finer grain understanding of the context of natural language.

For the purpose of the research the pre-trained Albert-base-v2 model from Hugging Face was implemented for Dataset 1 & 2. The Albert-base-v2 model had 12 repeating layers, 768 hidden layers and 11 million parameters. The models were run via the Simple Transformers library and the results tracked using the W&B website. The max sequence length for Dataset 1 was set at 150 and Dataset 2 at 235 as discussed above. The training process ran for 3 epochs but had an early stopping patience of 5. The learning rate was configured at 4e-5 with Adam optimization of 1e-8. During the training stage the training and development (dev) dataset were used with the dev dataset used to ensure the model didn't overfit. After the models had run the test dataset was used for predicting the final star rating.

4.5.2 Experiment 1: Evaluation and Results

The overall accuracy gained by Dataset 1 was 67.49% giving an error rate of 32.51%. During training a final MCC score of 58.62% and accuracy of 66.88% were achieved. The accuracy scores during validation and testing would indicate that the model did generalize well and did not overfit. Objective CI outlined in Table 1 has successfully been achieved.

4.5.3 Experiment 2: Evaluation and Results

For Dataset 2 a final accuracy score of 68.39% was achieved with an error rate of 31.61%. The MCC score during training was 60.61% with an accuracy score of 68.48%. This would indicate the model did generalize well and did not overfit. Objective CII outlined in Table 1 has successfully been achieved. Overall, the model deployed for experiment 2 achieved the highest accuracy score and the confusion matrix for the model is displayed below.

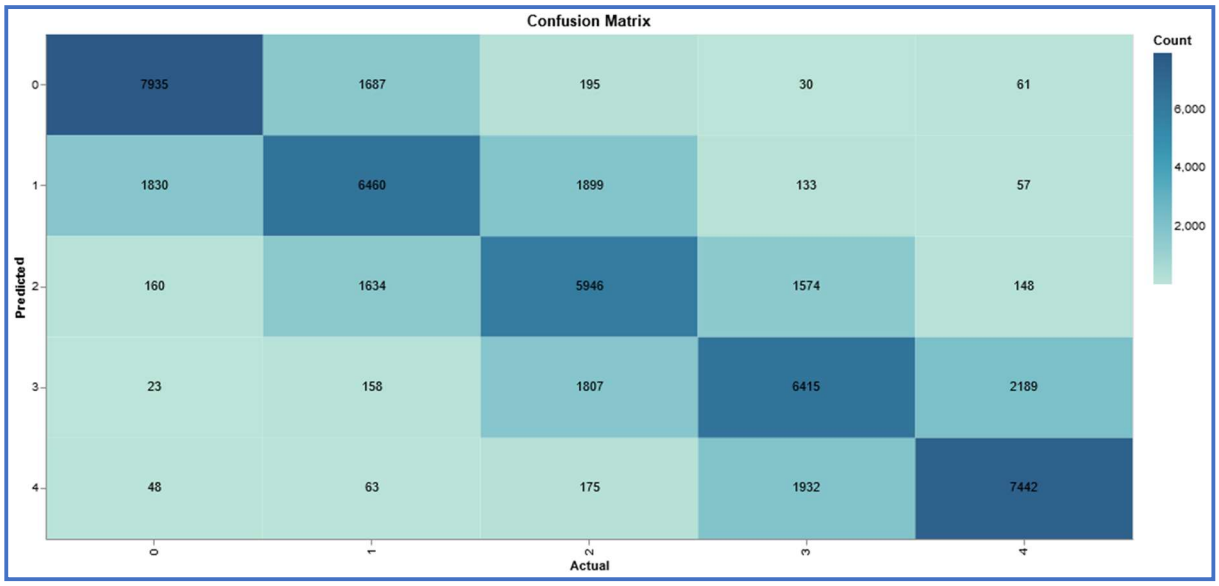

Figure 5: Confusion Matrix of ALBERT-base D2

Figure 5 shows the actual star rating on the X-axis and the predicted star rating on the Y-axis. The model is best at predicting very negative (0) followed by very positive (4) then negative (1) and positive (3) and finally the model finds it most difficult to predict neutral reviews (2). It is also clear from the confusion matrix that the model's highest misclassification rate is when it predicts either the preceding or following star of the actual outcome. The difference in prediction accuracy between 1-3 stars is not huge and this would tie with previous research by Zhu, Moh and Moh (2016). The authors found that star rating between 2-4 (1-3) were easily misclassified and the hardest to predict. In a lot of previous research this was one reason why these star ratings were grouped as neutral.

4.6 Implementation, Evaluation and Results of ELECTRA-base Models

4.6.1 Implementation

ELCTRA was developed by research teams at Google Brain and Stanford University and was developed to overcome the drawbacks of BERT's pre-training. The model was released with the paper ELECTRA: Pre-Training Text Encoders as Discriminators Rather Than Generators (Manning, 2020). The authors concluded that improvements could be achieved by concentrating on the objective of MLM. The researchers found that there was an efficiency limit with the extent that Tokens could learn an understanding of language context when only 15% of the tokens were masked. Furthermore, the masked tokens were only present when conducting the pre-training and not for fine-tuning resulting in a distribution of tokens that was different for each stage. As a result, the authors developed a model based on a generator and a discriminator. A generative model is a model that attempts to predict the word i.e. BERT with MLM while a discriminator attempts to predict a label or class.

The Generator model applies a small BERT model with MLM to construct a sequence made up of 15% masked tokens resulting in an incorrect series. Next the incorrect sequence which constitutes a series of replaced tokens and original tokens is fed into the Discriminator model. The Discriminator models attempts to predict if a token is an original token or a replaced token. The result was a model that learned from every token rather than just the 15% of masked token in the original BERT model. After the generator has done its job it's discarded, and the discriminator is used as the pre-trained model.

The ELECTRA-base discriminator model was the second model applied as part of the research. The model consists of 12 layers, 768 hidden layers and 110 million parameters. The configuration setting for the models were the same as the ALBERT-base as the candidate wanted to achieve a fair comparison of each model. After the model had finished the training process the test dataset was used for the prediction of star rating

4.6.2 Experiment 3: Evaluation and Results

The overall accuracy gained by Dataset 1 was 67.92% giving an error rate of 32.08%. During training a final MCC score of 59.72% and accuracy of 67.75% was achieved. This would indicate that the model did generalize well and did not overfit. Objective CIII outlined in Table 1 has successfully been achieved

4.6.3 Experiment 4: Evaluation and Results

For Dataset 2 a final accuracy score of 68.47% was achieved with an error rate of 31.78%. The final MCC score during training was 60.51% and an accuracy score of 68.55%. Again, this would indicate that the model did generalize well and did not overfit. Objective CIV outlined in Table 1 has successfully been achieved

4.7 Implementation, Evaluation and Results of Smaller BERT Models

4.7.1 Implementation

The Smaller BERT models were introduced by the research paper Well-Read Students Learn Better: On the Importance of Pre-training Compact Models (Turc et al., 2019). The goal of the research was to produce smaller models to run on restricted computational resources. The research garnered 24 models ranging from 4 million parameters to 110 million. The 24 models were based on Pre-trained Distillation. The process of pre-trained distillation was a series of three standard training objectives. First pre-training of an unlabelled language model data consisted of the training of a compact model with MLM that learned the context of language from a large text corpus. Next distillation occurred where the teacher model transferred its knowledge to the smaller student. Finally fine-tuning can be done on a labelled dataset.

For the purpose of this research the BERT-base which was included in the 24 models was not utilised as this had been released previously by Devlin et al. (2019). BERT-base was retrained for the purpose of completeness and included in the release of the smaller BERT models. The model utilised had 10 repeating layers, 768 hidden layers and 95.9 million parameters. The previous configuration was followed except for setting the argument of do_lower_case to true as the pre-trained model was uncased. Uncased refers to the text having been set to lowercase as part of tokenisation.

4.7.2 Experiment 5: Evaluation and Results

The overall accuracy gained by Dataset 1 was 67.19% giving an error rate of 32.81%. During training a final MCC score of 58.89% and accuracy of 67.1% was achieved. Objective CV outlined in Table 1 has successfully been achieved

4.7.3 Experiment 6: Evaluation and Results

For Dataset 2 a final accuracy score of 67.76% was achieved with an error rate of 32.24%. The final MCC score during training was 59.86% and an accuracy score of 67.88%. This would indicate again that the model did generalize well and did not overfit. Objective CVI outlined in Table 1 has successfully been achieve.

5 Discussion

5.1 Comparison of Developed Yelp Models

The comparison of the developed models was achieved by utilising the experiment tracking website W&B and visualising the data with Tableau. When tracking each model, the application calculated the metrics chosen by the user such as MCC and Accuracy but also generated tables such as a confusion matrix. As both datasets were balanced accuracy was chosen as the overall metric to decide on the best performing model. Accuracy was also chosen as it was necessary to calculate Error Rate for the comparison of the best model to previous research in the prediction of Yelp star rating. The final accuracy score for the 6 models is displayed in below diagram

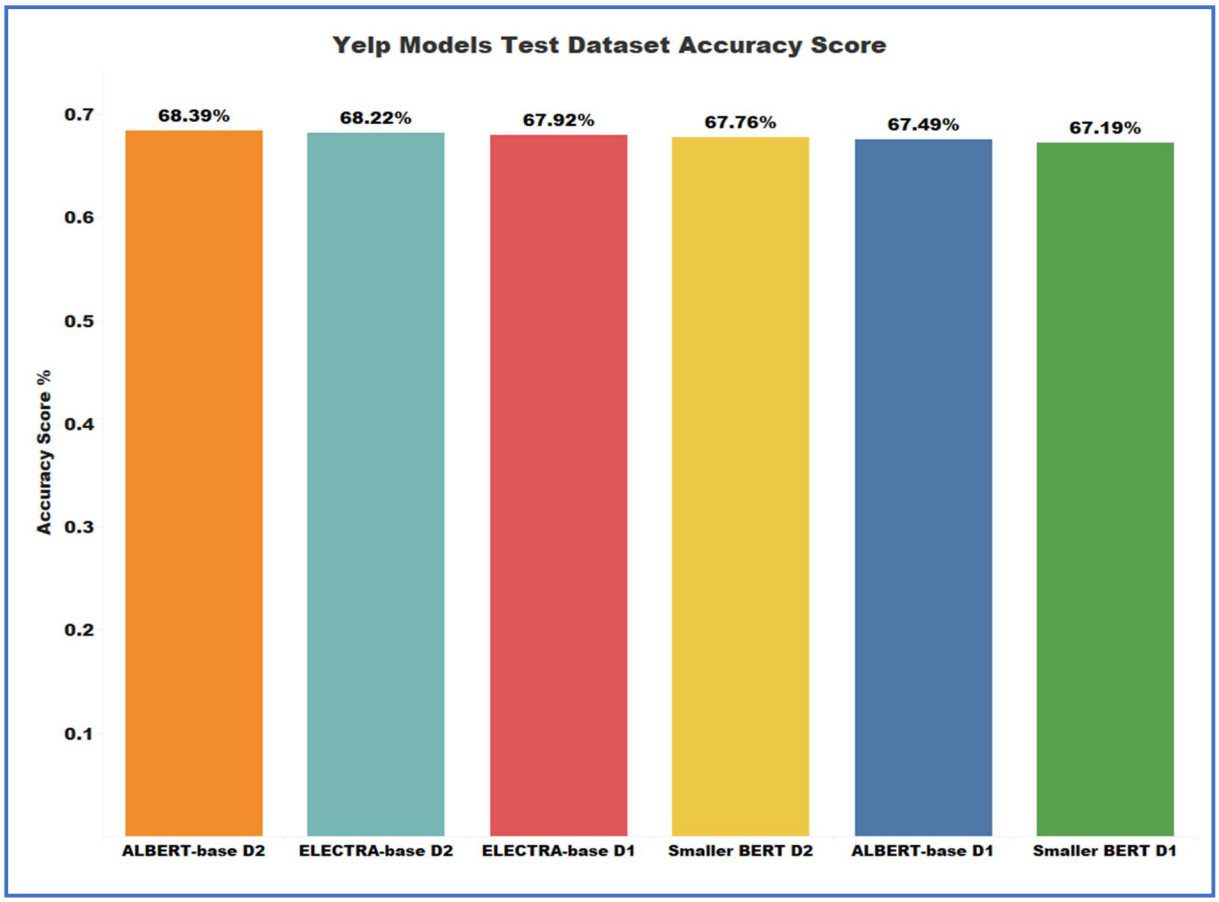

Figure 6: Final Test Dataset Accuracy Score

From Figure 5 ALBERT-base D2 achieved the highest accuracy with 68.39%. However, it was closely followed by ELECTRA-base D2 and ELECTRA-base D1 with 68.22% and 67.92%

respectively. All three models based on Dataset 2 performed comparably as the difference between ALBERT-base and Smaller BERT was only 0.63%. Importantly the three models based on dataset 2 outperformed the same models that used Dataset 1. The difference between each model for Dataset 1 compared to Dataset 2 ranged from Albert-base (0.9%), ELECTRA-base (0.3%) and Smaller BERT (0.57%). The difference between the highest and lowest scoring models was 1.20% (ALBERT-base D2 and Smaller Bert D1)

However, is the difference significant enough for the time it takes to run the models. When running these models with a cloud service there is a cost allocated so therefore users might have a concern with this. The time taken for training and testing each model can be seen in below diagram.

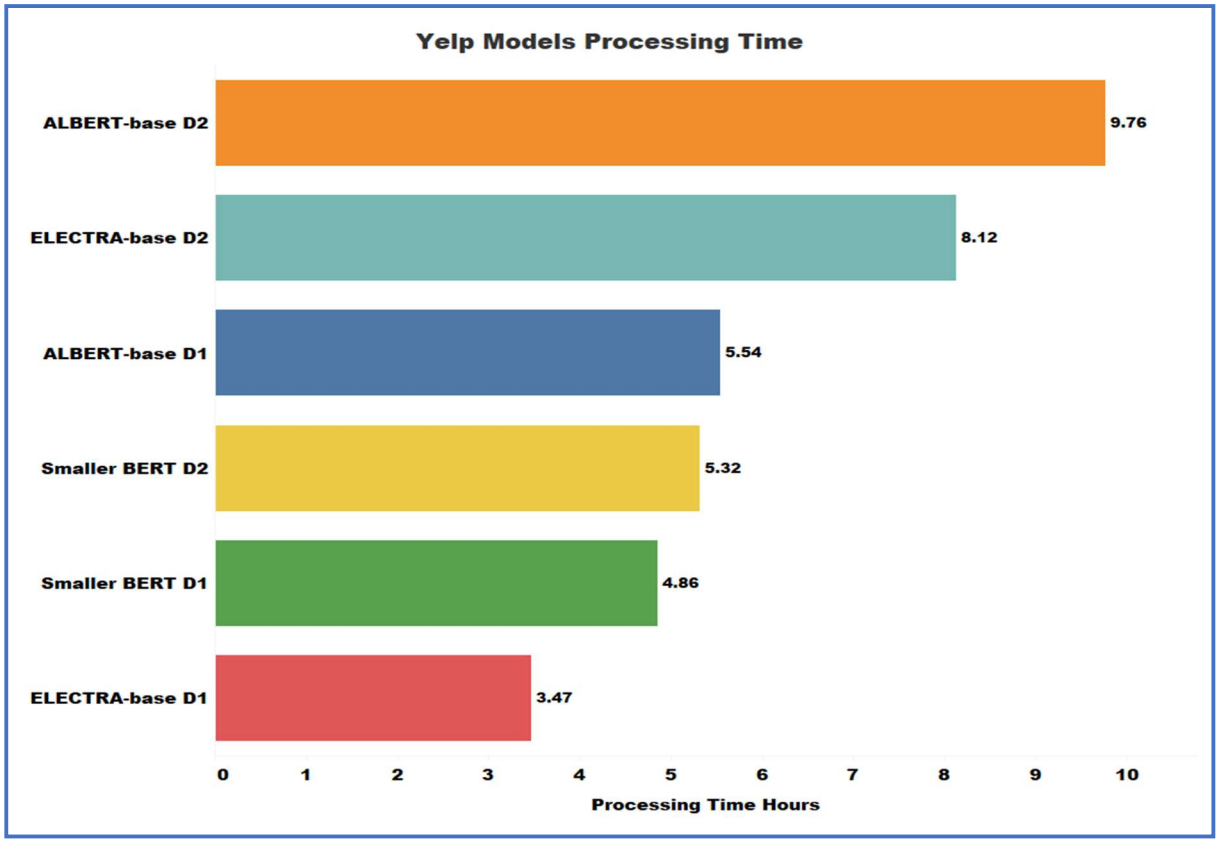

Figure 7: Yelp Models Processing Time

From Figure 6 the longest training time was for ALBERT-base D2 (9.76 hours) and the shortest ELECTRA-base D1 (3.47 hours). The results for ALBERT-base D2 was 68.39% the highest achieved and ELECTRA-base D1 67.92% which was the third highest. If a user similarly to the candidate had to use cloud GPU's to run the models and was concerned with the cost associated, then by sacrificing a slight drop in accuracy a user can still achieve comparable results using the quickest running model.

5.2 Comparison of Developed Models verses State of the Art Models

From Table 2 in the Literature Review the XLNet model developed by Yang et al. (2019) achieved the lowest Error Rate of 27.80% followed by BERT-Large+ITPT 28.62% developed by Sun et al. (2019). However, the model developed by the candidate was based on a light Bert which was developed to reduce the size of BERT while upholding comparable results.

Reducing the size of the BERT model, has had a knock-on effect of an increase in Error Rate. Therefore the candidates model is more comparable to FastBERT developed by Liu et al. (2020). SOTA results are achieved when a user has vast access to computational resources. The reasons for difference of Error Rate are presented below.

Author	Dataset	Model	Error Rate	Reason for Higher/Lower Score
(Liu et al., 2020)	Yelp 5-star	FastBERT	35.07%	The model due to reducing running time is not as powerful resulting in a higher Error Rate
(Sun et al., 2019)	Yelp 5-star	BERT-Large+ITPT	28.62%	The model is based on BERT-Large which requires considerably more computational resources to run resulting in a lower Error Rate
(Yang *et al.*, 2019)	Yelp 5-star	XLNet	27.80%	The model is based on XLNet-Large which requires considerably more computational resources to run resulting in a lower Error Rate

Figure 8: Comparison to Developed Models

6 Conclusion and Future Work

6.1 Conclusion

The objective of the research undertaken was the implementation of a fine-grained sentiment analysis of Yelp reviews using deep learning techniques when impacted by computational resources. Furthermore, the researched attempted to show if it was possible to increase the overall accuracy score of the models by selecting useful reviews.

Three models were selected to accomplish the objectives as they were specifically designed to overcome computational restrictions. The models selected were ALBERT-base, ELECTRA-base and Smaller BERT models. ALBERT-base was based on PR, ELECTRA-base on RTD while Smaller BERT was developed using KD. The models were run via the Simple Transformers library with the same configuration applied to give a fair comparison of each model.

ALBERT-base D2 achieved the highest accuracy score of 68.39% and showed that it was possible to achieve comparable accuracy scores on limited computational power. The research also showed that an increase in accuracy was gained by selecting useful reviews. The average increase was 1% so a user needs to decide if this justifies the time required to run the model.

6.2 Future Work

As the models were run on the same configuration for comparison reasons for future work W&B Sweeps would be utilised. Sweeps allows a user to run hyperparameter optimization on various combinations returning the optimal configuration to achieve highest accuracy. However due to time constraints the candidate was not able to be implement this and for future research this would be applied. The candidate achieved comparable results to previous research

however to achieve SOTA results access to more computational power is required. Future work would include running ALBERT-xxlarge and ELECTRA-large to accomplish better results.

6.3 Acknowledgements

I would like to specially acknowledge and offer my thanks to Dr. Catherine Mulwa for her guidance and support throughout the research period.

References

Azevedo, A. and Santos, M. F. (2008) 'KDD, semma and CRISP-DM: A parallel overview', *MCCSIS'08 - IADIS Multi Conference on Computer Science and Information Systems; Proceedings of Informatics 2008 and Data Mining 2008*, (June), pp. 182–185.

Baldi, P. *et al.* (2000) 'Assessing the accuracy of prediction algorithms for classification: An overview', *Bioinformatics*, 16(5), pp. 412–424. doi: 10.1093/bioinformatics/16.5.412.

Bongirwar, V. K. (2015) 'A Survey on Sentence Level Sentiment Analysis', *International Journal of Computer Science Trends and Technology*, 3(3), pp. 110–113. Available at: www.ijcstjournal.org.

Chaovalit, P. and Thou, L. (2005) 'Movie review mining: A comparison between supervised and unsupervised classification approaches', *Proceedings of the Annual Hawaii International Conference on System Sciences*, 00(C), p. 112. doi: 10.1109/hicss.2005.445.

Clark, K. *et al.* (2020) 'ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators', pp. 1–18. Available at: http://arxiv.org/abs/2003.10555.

Clement, J. (2019a) *Online reviews - Statistics & Facts*, *Statista*. Available at: https://www.statista.com/topics/4381/online-reviews/%0A.

Clement, J. (2019b) *Types of information internet users in the United States are likely to search on their smartphones instead of speaking to a store associate as of March 2019*, *Statista*. Available at: https://www.statista.com/statistics/1046672/information-us-internet-users-look-for-on-smartphone-vs-store-associate-2019/%0A.

Clement, J. (2020a) *Number of unique mobile visitors to Yelp from 3rd quarter 2016 to 4th quarter 2019*, *Statista*. Available at: https://www.statista.com/statistics/385440/unique-mobile-visitors-yelp/%0A.

Clement, J. (2020b) *Social media - Statistics & Facts*, *Statista*. Available at: https://www.statista.com/topics/1164/social-networks/#dossierSummary__chapter1%0A.

Clement, J. (2020c) *Yelp's net revenue from 1st quarter 2010 to 4th quarter 2019*, *Statista*. Available at: https://www.statista.com/statistics/278651/yelps-quarterly-net-revenue/%0A.

Devlin, J. *et al.* (2019) 'BERT: Pre-training of deep bidirectional transformers for language understanding', *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of*

the Conference, 1(Mlm), pp. 4171–4186.

Howard, J. and Ruder, S. (2018) 'Universal language model fine-tuning for text classification', *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1, pp. 328–339. doi: 10.18653/v1/p18-1031.

Hu, M. and Liu, B. (2004) 'Mining and summarizing customer reviews', *KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168–177. doi: 10.1145/1014052.1014073.

Johnson, R. and Zhang, T. (2017) 'Deep pyramid convolutional neural networks for text categorization', *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1, pp. 562–570. doi: 10.18653/v1/P17-1052.

Kemp, S. (2020) *Digital 2020*, *Datareportal*. Available at: https://datareportal.com/reports/digital-2020-global-digital-overview%0A.

Kim, H. and Arguello, J. (2017) 'Evaluation of features to predict the usefulness of online reviews', *Proceedings of the Association for Information Science and Technology*, 54(1), pp. 213–221. doi: 10.1002/pra2.2017.14505401024.

Kunst, A. (2020a) *How do you search for specific information on a product that you want to buy?*, *Statista*. Available at: https://www.statista.com/forecasts/997051/sources-of-information-about-products-in-the-us%0A.

Kunst, A. (2020b) *Which of these statements on online shopping do you agree with?*, *Statista*. Available at: https://www.statista.com/forecasts/997187/attitudes-towards-online-shopping-in-the-us%0A.

Lan, Z. *et al.* (2019) 'ALBERT: A Lite BERT for Self-supervised Learning of Language Representations', pp. 1–17. Available at: http://arxiv.org/abs/1909.11942.

Liu, B. (2012) *Sentiment analysis and opinion mining*, *Synthesis Lectures on Human Language Technologies*. doi: 10.2200/S00416ED1V01Y201204HLT016.

Liu, W. *et al.* (2020) 'FastBERT: a Self-distilling BERT with Adaptive Inference Time'. Available at: http://arxiv.org/abs/2004.02178.

Lu, Y. *et al.* (2010) 'Exploiting social context for review quality prediction', *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, pp. 691–700. doi: 10.1145/1772690.1772761.

Luca, M. (2012) 'Reviews, Reputation, and Revenue: The Case of Yelp.Com', *SSRN Electronic Journal*. doi: 10.2139/ssrn.1928601.

Manning, C. D. (2020) 'Electra : Pre - Training Text Encoders As Discriminators Rather Than Generators', *Iclr*, pp. 1–18. Available at: https://arxiv.org/pdf/2003.10555.pdf.
MLlib - Basic Statistics (no date) *Apache*. Available at: https://spark.apache.org/docs/1.2.0/mllib-statistics.html.

Moh, M. *et al.* (2016) 'On multi-tier sentiment analysis using supervised machine learning', *Proceedings - 2015 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2015*. IEEE, 1, pp. 341–344. doi: 10.1109/WI-IAT.2015.154.

Munikar, M., Shakya, S. and Shrestha, A. (2019) 'Fine-grained Sentiment Classification using BERT', *International Conference on Artificial Intelligence for Transforming Business and Society, AITB 2019*, pp. 2–5. doi: 10.1109/AITB48515.2019.8947435.

Pan, S. J. and Yang, Q. (2010) 'A survey on transfer learning', *IEEE Transactions on Knowledge and Data Engineering*. IEEE, 22(10), pp. 1345–1359. doi: 10.1109/TKDE.2009.191.

Pang, B. and Lee, L. (2005) 'Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales', *ACL-05 - 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, (1), pp. 115–124.

Pang, B., Lee, L. and Vaithyanathan, S. (2002) 'Thumbs up? Sentiment Classification using Machine Learning Techniques', (July), pp. 79–86. Available at: http://arxiv.org/abs/cs/0205070.

Pathak, A. R. *et al.* (2020) *Application of Deep Learning Approaches for Sentiment Analysis*, *Algorithms for Intelligent Systems*. Singapore: Springer. doi: 10.1007/978-981-15-1216-2_1.

Price, M. (2011) *Tapping our powers of persuasion*, *American Pyscholoigical Association*. Available at: https://www.apa.org/monitor/2011/02/persuasion.

Sun, C. *et al.* (2019) 'How to Fine-Tune BERT for Text Classification?', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11856 LNAI(2), pp. 194–206. doi: 10.1007/978-3-030-32381-3_16.

Taboada, M. *et al.* (2011) 'Lexicon-based methods for sentiment analysis', *Computational Linguistics*, 37(2), pp. 267–307. doi: 10.1162/COLI_a_00049.

Turc, I. *et al.* (2019) 'Well-Read Students Learn Better: On the Importance of Pre-training Compact Models', (Mlm), pp. 1–13. Available at: http://arxiv.org/abs/1908.08962.

Turney, P. D. (2002) 'Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews', (July), pp. 417–424. Available at: http://arxiv.org/abs/cs/0212032.

Vijayan, V. K., Bindu, K. R. and Parameswaran, L. (2017) 'A comprehensive study of text classification algorithms', *2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017*, 2017-January, pp. 1109–1113. doi: 10.1109/ICACCI.2017.8125990.

Yang, Z. *et al.* (2019) 'XLNet: Generalized Autoregressive Pretraining for Language Understanding', (NeurIPS), pp. 1–18. Available at: http://arxiv.org/abs/1906.08237.

Zhang, X., Zhao, J. and Lecun, Y. (2015) 'Character-level convolutional networks for text

classification', *Advances in Neural Information Processing Systems*, 2015-Janua, pp. 649–657.

Zhu, Y., Moh, M. and Moh, T. S. (2016) 'Multi-layer text classification with voting for consumer reviews', *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016*. IEEE, pp. 1991–1999. doi: 10.1109/BigData.2016.7840821.

Zhuang, F. *et al.* (2019) 'A Comprehensive Survey on Transfer Learning', pp. 1–31. Available at: http://arxiv.org/abs/1911.02685.