Master Thesis

(Msc. Data Science and Marketing Analytics 2020 – 2021)

ERASMUS UNIVERSITEIT ROTTERDAM

Say Hello to the Future of Beauty:

Evaluation of Virtual Try-On Technology for Cosmetic Products Using Text Analytics

Author:
Kenita Hadi
Student Number:
535977

Supervisor:
Dr. V Avagyan
Second Assessor:
Dr. A Tetereva

September 17, 2021

Abstract

Immersive technologies, such as Augmented Reality (AR) driven Virtual Try-On technology (VTO) have recently attracted investors and managers on their adoption, since they help in facilitating the transition of physical store to ecommerce through virtual product testing (Hopping, 2020). Past research has shown that the presence of AR and VR improve purchase decision, but their methodologies are heavily relied on experiment. This paper, therefore, presents a deeper analysis from the current beauty VTO apps' using the readily available review data of users regarding the features of the technology that matter the most in determining its rating score. We dive into various Natural Language Processing techniques that allow both theory-based of Technology Acceptance Model (TAM) and thematic features extraction (LDA). We find that though the quantitative performance of thematic feature extraction slightly outperforms TAM based features, the interpretability of concept-based features is still preferred. Moreover, from these models we found that users' sentiment, ease of use, visuals quality (immersion and vividness) and interactivity of the apps are among strongest determinant of the VTO's rating.

Acronyms

AR - Augmented Reality

VR – Virtual Reality

VTO - Virtual Try On

TAM - Technology Acceptance Model

PU - Perceived Usefulness

PEOU - Perceived Ease of Use

ATU – Attitudes Towards Using

BI – Behavioral Intention

PI – Perceived Interactivity

WOM - Word-of-Mouth

NLP - Natural Language Processing

ML – Machine Learning

LDA - Latent Dirichlet Allocation

Table of Contents

1. Introduction

Over the recent decade, new technologies have allowed the modification of the retail landscape (Hopping, 2000). The global retail sector that is gradually transitioning from multi-channel to omnichannel recognizes the integration of physical store, mobile and internet to sell products and services (Park and Yoo, 2020). Moreover, the COVID-19 pandemic amplifies the greater significance of online retailing, forcing businesses to leave the conventional business model and shifts their strategies by utilization of technologies.

To further accommodate the integration that mimics the sensory and convenient experience of offline shopping, the modern retail industry is now characterized by integrating technologies that explore immersive technologies such as Mixed Reality (MR), Virtual Reality (VR) and Augmented Reality (AR). One of the latest and emerging developments, Augmented Reality (AR) in particular, is experiencing a massive popularity among investors, companies and consumers. The market size of AR that generated over $640.2 million of revenue in 2015 is forecasted to grow substantially at a CAGR by nearly 75% in 2025, valued at $35 billion dollars (Goldman Sachs, 2016). Furthermore, the adoption rate is expected to be comparable to that of smartphones by then. Today, a growing number of firms have embedded AR as part of their omnichannel strategy.

The pandemic has especially raised awareness on personal safety and hygiene. Thus, the beauty sector, comprising skin care, cosmetics, hair care, and personal care is one of the industries that is highly impacted by the pandemic. Their outlets all over the world are forced to close temporarily to limit physical contacts. In response to the pandemic and greater needs in incorporating new technologies to facilitate e-commerce, leading beauty companies such as L'oreal and Maybelline have recently integrated AR to produce virtual makeup try-on tools, naming Modiface and Virtual Try-On (VTO) in 2018 that allow customers to apply and compare various makeup products. YouCam, a Taiwanese makeup try-on application experienced an increase of 32% of downloads since the start of the pandemic (Digiday, 2021). Additionally, Sephora's Virtual Artist enables features that examine an individual's skin type to recommend the appropriate products to try on, providing customers with a personalized and unique online shopping experience.

This study focuses on the role of AR virtual try-on for cosmetics products as marketing tools in smart devices, which are among the most common types of adoption in retail businesses (Javornik, 2014). AR VTO has become a mainstream and easily available technology on smart devices, equipped with operating systems, interactive screens, stimuli recorder such as cameras, speakers and location-based sensors. AR VTO can enhance the online shopping experience by stimulating mental image of the visually perceived objects. Consequently, it improves consumers' understanding about the products and in making a reliable purchase decision without having to go to the physical stores. Some popular AR VTO applications include IKEA place, which allows customers to place virtual furniture in their room. Others in fashion retails that allow users to try products virtually include eyewear stores such as Ray-ban and Ace & Tate, footwear brands such as Nike, Adidas, and Gucci, and clothing brands such as H&M is planning to include a digital fitting room by summer 2021.

Although it appears that retail practitioners are surging more attention to AR, less than 35% of them have adopted this technology (Park and Yoo, 2020) due to the limited knowledge on its effectiveness in generating positive business outcomes. Marketeers are still uncertain whether AR serves the purpose to solely entertain users (Owyang, 2010). A survey conducted by DigitalBridge (2017), shows that 51% of their respondents think that firms are not able to utilize AR to its full potential. Some AR try-on applications such as JC Penney, Converse and Tobi Fashion do not exist anymore (Accenture, 2014) that is potentially due to the low technology acceptance by users and poor design that failed to meet customer base demands.

This research aims to bridge these knowledge gaps in three folds; industry, source of data and methods. Firstly, we want to investigate the important aspects of VTO technology for cosmetic products given that research incorporating AR in the beauty industry is scarce; the first Virtual Try-On was only introduced in 2018 by L'oreal. Recent research incorporating AR in the online retail are found mainly in the fashion (Bonnin, 2020, Vasquez-Parraga et al., 2017), gaming (Liarokapis, 2006, Pallavicini et al., 2019), and furniture industry (Ozturkcan, 2020, Rauschnabe et al., 2019). It remains uncertain whether these findings can be generalized to the beauty industry, due to the nature of different types of products, target customers, and purchase intentions. Moreover, we would like to investigate the effect of two supporting variables that are applicable to this research, naming review time and hedonic/utilitarian aspect of the VTO technology. Review time refers to the pre and post corona time that allows us to investigate its expected relationship with users' intention to use; whether they are more encouraged to use the apps due to corona. Hedonic and utilitarian aspects of VTO technology allows the assessment of their importance, this provides insight whether VTO is currently perceived more as an entertainment or functional tools. Thus, by providing these relevant attributes we hope to guide managers in successful VTO implementation and aspire the beauty retail industry to adopt VTO technology to its full potential.

Secondly, we want to contribute to the academic practice by exploiting the abundance and readily available online textual data to generate marketing insights. Given that numerous scholars have studied the effectiveness of AR as an ecommerce tool, their methods are limited to experiment and surveys in a controlled setting with a large focus on measuring behavioral intention (Poushneh et al., 2016, Sauer et al., 2017, Beck and Crie, 2018). While they consider the representativeness of the samples, they might not be sufficient to describe all users. The low technology acceptance entails the "unfamiliarity" of the potential users and the lack of knowledge on product design by managers, that signifies the importance of information and opinions exchange by users that have experienced AR; also referred to word-of-mouth (WOM) communication (Rese et al., 2014). This mode of communication these days rely heavily on electronic WOM or online reviews/ratings that have become a trusted source in driving purchase decisions (Chevalier and Mayzlin, 2006) that will be the data source of this research. Thus, to capture the real time sentiment and eWOM of the current users, a textual review data from one of the best rated and most downloaded makeup editor mobile apps, YouCam is utilized for this research.

Lastly, we are interested in the "*how*" to process this text data by exploration of various text analysis techniques that allow incorporation of concepts and statistical approaches, with a supplement of machine learning techniques that provide the best balance between complexity and interpretability. In particular, we review both unsupervised and semi supervised text extraction methods to assess "*what are the current users talking about?*" and "*how do they feel about the apps and its attributes?* The field of semi-supervised NLP is rather new yet offers a great usability and flexibility to complement researcher's knowledge about the topic and undetected pattern of the data. Text data, in combination with the researchers' ability have been proven useful to extract underlying insight, in tracking, measuring, understanding and interpreting causes and consequences of marketplace behavior (Berger et al., 2019). We hope to capture this opportunity that broadens the application of text analysis in the marketing academia field.

Altogether, this research aims to provide clarity to these problem statements and delivering the contributions by answering the following research questions:

RQ1: *How can text analysis method be applied in AR-driven YouCam Makeup apps' reviews to predict online rating?*
RQ2: *What are the most important drivers of the AR-driven YouCam Makeup apps in predicting online rating?*

Throughout this paper, we will answer these questions by exploring the current literatures in immersive technology (Chapter 2), then apply the general procedure of text analysis for marketing insights in Chapter 3 and Chapter 4 from data preprocess, feature extraction and predictive analysis (Berger, et. al, 2019). The final analysis in Chapter 5 will use the extracted features as input for our predictive models to the rating score of YouCam Makeup apps and the interpretation of our findings will be based on the best performing model.

2. Literature Review

This part reviews the current literatures regarding the usage of AR technology by introducing the definition and usability of AR Virtual Try-On for e-commerce purposes, related research and Technology Acceptance Model (TAM) to gain the first impression on relevant AR VTO attributes that can be extracted from YouCam review data.

2.1 Overview of AR Virtual Try-On in E-commerce

There are three major types of immersive technology; Virtual Reality (VR), Augmented Reality (AR) and Mixed Reality (MR) which combines VR and AR. The major distinction between VR and AR is that VR generates a perception of reality by stimulating senses (vision, hearing, touch) based on solely virtual information (Wedel et al., 2020), while AR allows capturing "a real time computer generated virtual imagery that imitates the real objects applied in the real setting" (Zhou et al., 2015) which is the focus of this research. The augmented objects may depend on the context of use companies are trying to augment; products and places are among the primary augmented objects for marketing purposes. For cosmetic products, it facilitates try-on simulations on live video that tracks facial features with high precision, resulting in a photo-realistic makeup simulation. This often also comes in a form of "magic mirror" that is placed in the actual retail store, which elevates the actual image of visitors with makeup. VTO technology is a form of marker-based AR, which requires unique marker or a static image to recognizes and triggers the augmented visuals. For the YouCam app, these triggers are facial features such as eyebrows, lips, hair and skin.

This powerful feature of AR makes it possible for brands to expand omnichannel experiences across customer journeys through an intuitive, context-sensitive and social connectivity interface (Hilken et al., 2018); customers are able to experience the "sensory richness" of testing the physical products in the form of imagery or animations. It is therefore especially relevant for digital shopping, more than 88% of potential buyers look for information online before deciding to purchase a product (Digitas Study, Vivaki advance, 2013). Thus, by delivering an enhanced interactive experience through visuals, usability and enjoyment, it may facilitate consumers' understanding about the products in making a reliable purchase decision and an improved experience without having to go to the physical stores. The effectiveness of AR as a marketing tool is determined by how successful it is in enhancing customer tasks and resemble shopping experience that generates the same desirable shopping outcomes in the real-world setting (Wedel et al., 2020), which requires AR to be highly accurate in its features of face and hair recognition, color customization, and face expression detection in our application of cosmetic products.

2.2 Related Work

Due to the broad application of immersive technology in many fields such as education, engineering, and psychology, research has attempted to measure the effect of AR presence to outcomes such as learning effectiveness, learning engagement and task performance[1]. Most studies that have investigated the effectiveness of AR in the marketing domain focus on measuring experiential value and behavioral intention to use immersive technologies of virtual reality (VR) and augmented reality (AR), using predictive metrics such as interactivity, vividness, and immersion and found that they positively influence pre-purchase intention; shopping enjoyment, online visits intention (Poushneh et al., 2016, Sauer et al., 2017), brand attitude and patronage intention (Park and Yoo, 2020, Beck and Crie, 2018) than a web-based presentation. For instance, Sauer et al., (2017) on their findings in AR integration in the fashion industry, indicate that AR provides an effective communication that results in positive attitude toward purchase intention, by generating a sense of immersion, enjoyment, usefulness and a greater novelty compared to web-based products presentation. Moreover, Wedel et al (2020) shows that a virtual imagery on products is useful in educating customers to make a reliable decision. Thus, it can be concluded that an effective Virtual Try-On application benefits customer by providing them an improved and convenient shopping experience that is personalized to their needs. Triggering similar responses on AR for cosmetic products may encourage more adoptions that positively reshape the future beauty shopping experience.

In term of methodologies, all these studies conduct experiments and surveys. We could only find a paper by Rese, et. al (2014) that performed a text analysis on online reviews of the VR driven Ikea Place apps and found that online reviews can replace surveys. They encourage future research to take advantage of the readily available online text data, that is more affordable and less time consuming than experiments. Furthermore, most recent research incorporating AR in the online retail environment are found mainly in the fashion (Bonnin, 2020, Vasquez-Parraga et al., 2017), gaming (Liarokapis, 2006, Pallavicini et al., 2019), and furniture industry (Ozturkcan, 2020, Rauschnabe et al., 2019). Research on AR in the beauty industry is therefore limited, the first Virtual Try-On was only introduced in 2018 by L'oreal. It remains uncertain whether these findings can be generalized to the beauty industry, due to the nature of different

[1] Refer to appendix

types of products, target customers, and purchase intentions. For instance, aspects such as product safety and hygiene hold great consideration in beauty product testing, it accelerates the needs of AR relative to other industries.

Table 1: *Overview of Research in AR/VR Marketing Applications from the Top Journals*

Source	Augmented Product	Key Findings	Variables
Bonnin (2020)	Ray Ban eyewear & Converse shoes	a.) AR presence does not increase patronage intention by utilitarian value but increases by hedonic value. b.) Familiarity with AR decreases AR perceived risk and increase patronage intention	Xs: Perceived risk of buying products online, utilitarian and hedonic evaluation, attractiveness of the online store, and the familiarity with AR Y: Patronage intention
Heller et al. (2020)	Desserts	AR presence improves decision comfort, motivate positive WOM and facilitates choice of higher value products in retail frontline, meditated by an improved processing fluency.	Xs: Mental imagery aspects, customers' processing type and fluency, product context Y: Behavioral intentions: Choice comfort and WOM
Park and Yoo (2020)	Youcam makeup products	The controllability and playfulness dimensions of perceived interactivity (PI) and mental imagery improve customers' attitude towards the product and consequently behavioral intention	Xs: Perceived interactivity, mental imagery, attitudes Y: Behavioral intention (willingness to purchase, willingness to revisit store)
Poushnesh (2018)	Night Sky, Cimagine furniture, and Ray Ban eyewear	a.) Consumers pay attention on both their privacy information and augmentation quality. b.) The ability to control access to personal information significantly affects users' satisfaction	Xs: Augmented quality and users' control of access to personal information Y: Users satisfactory experiences with AR
Rauschnabel et al. (2019)	IKEA furniture and Tunnel (projects song-related information to user's environment)	Consumer inspiration provision by AR meditates the perceived benefit from AR and changes in brand attitude	Xs: Augmentation quality, inspiration, utilitarian and hedonic benefits Y: Changes in brand attitude

Sauer et al. (2017)	Sunglasses and watches	a.) AR activates positive attitude towards medium and purchase intention by providing effective communication benefits compared to web-based product b.) Immersion meditates the relationship between interactivity and usefulness & enjoyment	Xs: Interactivity, vividness, immersion, media enjoyment, usefulness, attitude towards AR Y: Purchase intention.

2.3 Conceptual Framework

Technology Acceptance Model (TAM)

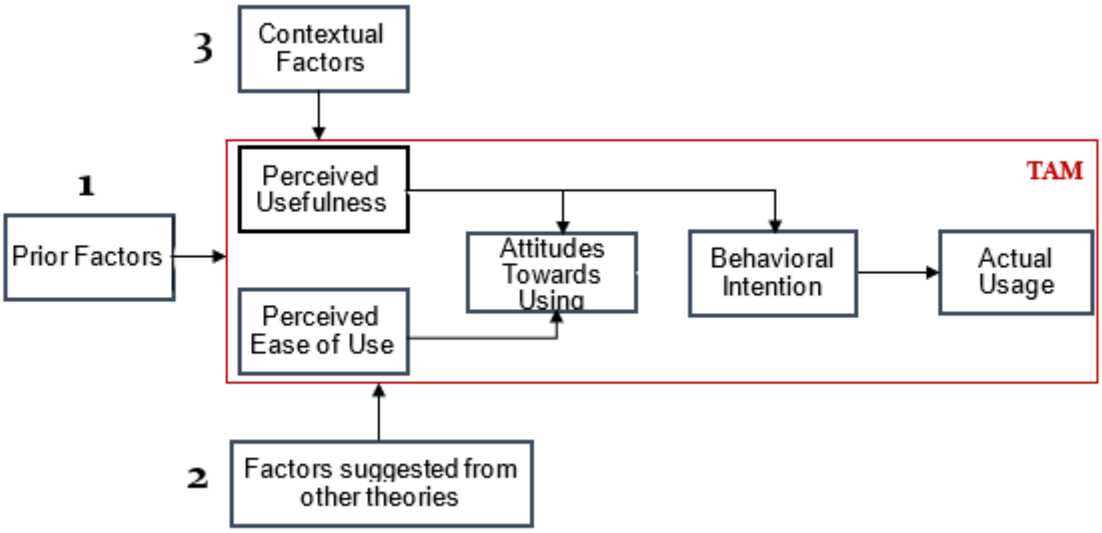
Though there exist a lot of theoretical foundations in AR research that vary from (but not limited to) conceptual blending theory, flow theory, and media richness theory, we decided to use the most popular framework of TAM due to its flexibility in allowing the inclusion of many variables, its stability and the model's validity in text mining application Rese et al (2014). Moreover, it could help in explaining the potential failure in adopting a technological innovation such AR resulted from the low technology acceptance or "unfamiliarity" of the potential users and the lack of a "knowledge structure" for product evaluation (Robertson and Gatignon, 1986). Due to the rapid development of new technologies over the decades, scholars have developed various frameworks and approaches to address this issue (King and He, 2006).

Technology Acceptance Model (TAM) originally developed by Davis in 1989 is one of the most widely used frameworks in Information Systems due to its simplicity and understandability. Moreover, King and He (2006) has confirmed the robustness of the model, the positive relationships among the five constructs in predicting actual usage and shown that it also offers a great applicability in various fields. TAM aims to explain the **behavioral intention (BI)** of users in adopting a new technology, with 2 primary extrinsic motivations: **perceived usefulness (PU)** and **perceived ease of use (PEOU)**, meditated by **attitudes (AT)** that contribute to the core aspects of TAM. In addition, Davis (1989) proposed the indirect positive effect of PEOU and PU.

Furthermore, TAM has been repetitively used and proved to be valid in explaining adoption of immersive technology such as Ikea Place and Ray-Ban VTO (Bonnin. 2020, Rese, et al., 2014, Poushned & Vasquez-Prrage, 2016). In the text mining context, Rese, et.al (2014) shows that online reviews are suitable to replace surveys on measuring the acceptance of technological innovation. For these reasons, TAM is chosen as the basis of conceptual theory that provides the first impression of dimensions that could help predict rating score of YouCam Makeup apps.

The usage of TAM as an instrument for meta-analysis has been applied in many empirical studies (Venkatesh and Davis, 2000), by allowing the extension and modifications of external variables influencing PU and PEOU. King and He (2006) summarize the 3 type of modifications as follows, [1] Prior factors include external factors regarding user's situational involvement such as experience in prior usage and personal self-efficacy, [2] factors such as expectation, subjective norm, task-technology fit, risk, and trust that are suggested by other theories to achieve higher predictive power of TAM; [3] contextual factors include the characteristic of users and the technology as suggested by the following figure:

Figure 1: *Tam and categories of modification (King and He, 2006)*



In addition, few studies have modified the consequence measures of actual usage to perceptual usage (Horton et al., 2001, Moon et al., 2001, Szajna et al., 1996). For the purpose of this analysis, we replace actual usage by **rating score** that can be generalized as users' perceived quality in using the innovative technology**.** Rating score is considered more appropriate as the dependent measurement for this dataset mainly due to the nature of the review data that implies all users have used the systems. Furthermore, we would like to incorporate more factors to generate more variables that may hold important roles in determining product rating and potentially improve the predictive model to the rating score (King and He, 2006). Due to the absence of data about the users', we can only incorporate other factors suggested from other immersive technology theories [2] to our extended TAM. Therefore, we select some of the most recurring and significance (to behavior intention) variables from past research in table 1 to complement our TAM variables. The following provides the definitions of these variables that will be used for our conceptual analysis:

***Table 2:** Overview of the Extended TAM Variables used for our Conceptual Analysis*[2]

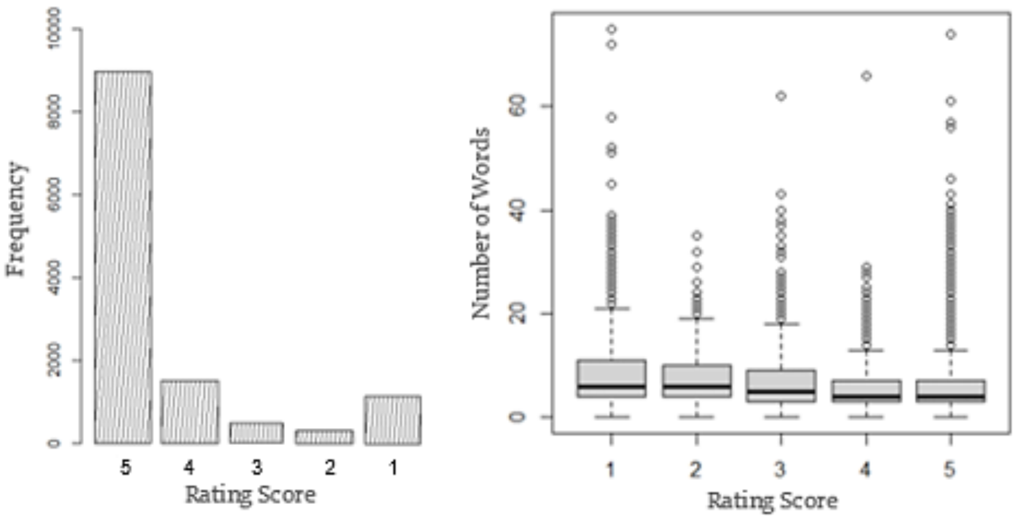
TAM core variables (King and He, 2006)	**Variables suggested from other theories**
Perceived Ease of Use (X1): The degree to which an individual perceives a (AR) technology as exhausting his or her cognitive resources.	**Perceived Interactivity (X5): T**echnological features that allow user control and website characteristics. It facilitates consumers' direct manipulation of objects on websites (Huang et al., 2010, Park & Yoo, 2020).
Perceived Usefulness (X2): The degree to which AR enhances contextual knowledge about the product, provides inspiration and facilitates decision making[3].	**Immersion (X6):** The degree of realism or user's authentic experience of the AR generated image to mimic the sensory experience, also referred as representational fidelity (Merchant et al., 2014, Park & Yoo, 2020).
Attitudes Towards Using (X3): User's judgment or affective responses of the desirability of using a new technological application.	**Vividness/media richness (X7)***:* Breadth and depth of the message: breadth being the number of sensory dimensions, cues, and senses presented (colors, graphics, etc.), and depth being the quality and resolution of the presentation (bandwidth). (Steuer, 1992, Huang & Liu, 2014)
Behavioral Intention (X4): User's willingness to use a specific technology (Park & Yoo, 2020).	**Enjoyment (X8):** The hedonic aspect of AR, an extent to which using a (AR) technology is perceived to be enjoyable for its own sake, without considering performance related outcomes (Rauschnabel et al, 2019, Sauer et al., 2017).
	Trust (X9): The expectation that the system application behaves in an ethical manner and does not take undue advantage, (ie. abuse of personal data) of a dependence upon them (Bonnin, 2020, Poushnesh, 2018).

[2] Refer to appendix A1, table 7
[3] Refer to additional references in appendix A1 table 8

3. Data

The data comprises 12440 reviews of "YouCam Makeup" mobile apps from its first release in December 2018 up to June 2021, scraped using Python from both US Play Store (https://play.google.com/store/apps/details?id=com.cyberlink.youperfect&hl=en_US&gl=US) and US Apple Store (https://apps.apple.com/us/app/you cam-makeup-selfie-editor/id863844475) with the average ratings of 4.5 and 4.8 respectively. YouCam Makeup allows users to try hundreds of cosmetic products in real time. The app is developed by a Taiwanese developer, Perfect Corp, that specializes in beauty tech solutions such as AI for skin diagnosis, skin shades detector, 3D try-on for makeup, jewelry and glasses (Perfectcorp, 2020). They have helped leading beauty enterprises all around the world such as Ardell, Benefit Cosmetic, Estée Lauder and Mac to incorporate these technologies as their ecommerce tools. YouCam, their best rated Apps, has won several awards including 2020 AI Excellence Awards in skin diagnostic technology, 2019 Beauty Innovation Awards, and 2020 BIG Innovation Awards (Perfectcorp, 2020). The scraped data consists of author, title, review text, review date and rating. We focus on the textual review data and exclude the author from the analysis. The reviews allow the evaluation of attributes of the apps that users favor or disfavor, and extraction of their opinions. Furthermore, the renowned reputation of the apps and the abundance of available online reviews motivate the decision to use this data for the purpose of this study. A glimpse at the summary statistics of the combined rating scores from Apple and Play Store are provided as follows:

Figure 2: *Frequency of rating scores (left) and document length on each rating class (right)*



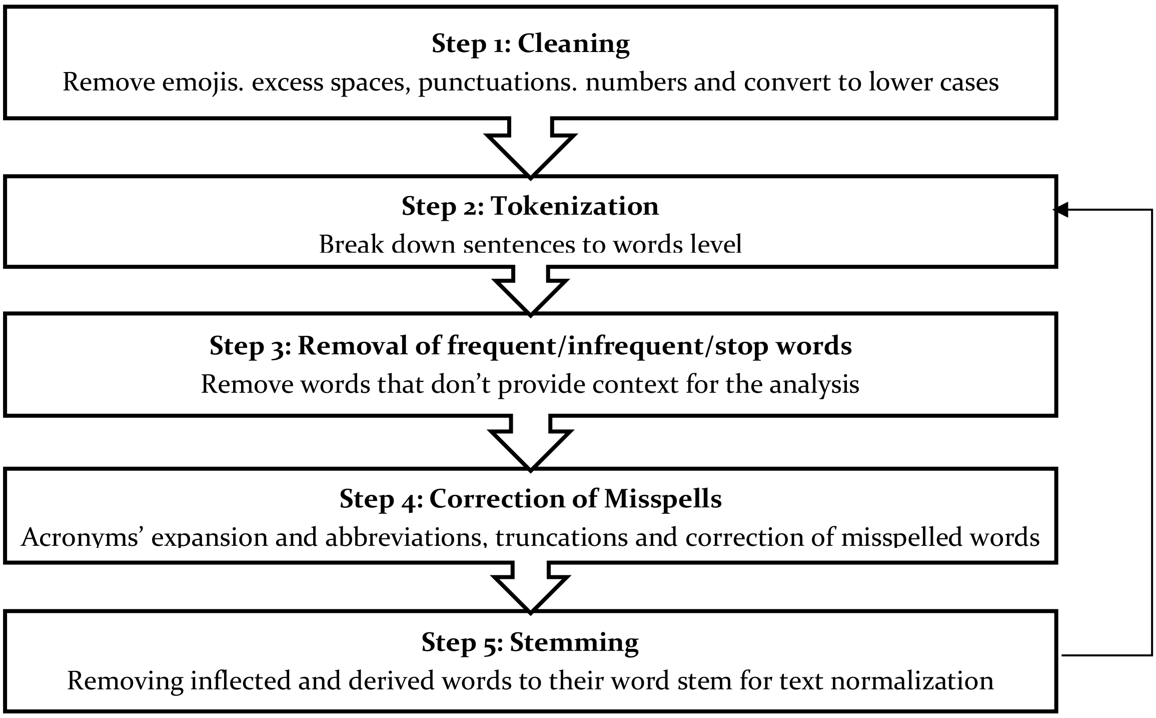
(a) **(b)**

The distribution of ratings (a) is highly imbalanced, with most users rated 5 starts. This may distort the predictive power of the classification model that is bias towards the majority class. Moreover, from (b), all reviews contain 6 words on average with no noticeable differences in each class and there are a few outliers with reviews that contain over 60 words that may destabilize the model. The selection of the predictive model will be carefully assessed to minimize these problems.

3.1 Data Preprocess

Text data, especially online reviews by nature are highly unstructured and large in features, thus require a lot of pre-cleaning before it can be processed as an input of any scientific analysis (Berger, et. al, 2019). For instance, this review contains misspelled words, emojis, or internet slang that do not exist in the English dictionary. NLP technique often uses a bag of N-grams and document term matrix that capture the co-occurrences of paired words (Kwartler, 2017). As the size of vocabularies increases so does the vector representation of documents, resulting in sparse vectors that take up high memory and computational power to run a text algorithm. Therefore, data preprocessing is the necessary first step as an attempt to reduce the number of vocabularies to generate features that provide inherent meaning to the context of this research, which are words that define AR features and opinions regarding the YouCam apps.

Figure 3: *Text pre-process procedure*


Step 1: Cleaning
Remove emojis. excess spaces, punctuations. numbers and convert to lower cases
Step 2: Tokenization
Break down sentences to words level
Step 3: Removal of frequent/infrequent/stop words
Remove words that don’t provide context for the analysis
Step 4: Correction of Misspells
Acronyms’ expansion and abbreviations, truncations and correction of misspelled words
Step 5: Stemming
Removing inflected and derived words to their word stem for text normalization

The preprocess follows the standard procedure of text mining that includes cleaning, tokenization, removing stop words and stemming (Berger, et. al, 2019, Kwartler, 2017). The first step of data pre-process is to remove emojis, excess spaces, punctuations, numbers, and upper cases using regular expression function: *gsub* from R for pattern matching and replacements. Afterwards, the sentences are broken down into word level through tokenization. Through this process, the frequency of word occurrences can be captured. To further narrow down the list of words, most frequent words of "app", "makeup" and their plural forms are removed since they are meaningless for features and sentiment extraction. Furthermore, stop words such as the, a, so, because that don't provide context are removed, resulting in 7288 unique words. The stop words library from tm package in R consists of a stop words dictionary that uses SMART and snowball lexicon[4]. There are exceptions to this removal:

a.) **Amplifier words:** these are adverbs that may change the intensity of adjectives or verbs such as "very, really, highly, more, extremely". They are kept since they will affect polarity scores in the sentiment analysis (ie."very good" >"good")

b.) **Negation Words:** words that provide opposite meaning to the words of sentences such as "don't, isn't, didn't, can't, not, wasn't". They are kept for the same reason with (a), they are also indicative to polarity score and provides opposing context in bi-grams extraction (i.e., "not useful" < "useful" and "isn't working" < "working")

c.) **Sentiment Words:** The SMART stop words lexicon contains words such as "useful, important, help" that will affect sentiments.

d.) **TAM indicator words:** Words that can provide context or be used as keywords to signal the TAM features are kept, these are words such as "changes" (perceived_interactivity), "accordingly" and "different" (immersion).

Subsequently, we check and correct for misspellings which is an optional task for text mining. Spell correction is highly applicable for our case since review data involves a cognitive process that may lead to typographical error. This type of error occurs in typesetting, both consistent (authors are either unsure about the correct spelling or think that they are using the right spelling) and conventional manners (authors intentionally misspell words due to haste) (Brown, 1988). Since text mining method often emphasizes words' importance based on their frequency, misspells correction can boost the accuracy of the prediction by recovering informative words (Yazdani et al., 2020), in this case they are valence shifters, sentiments, product attributes,

[4] https://cran.r-project.org/web/packages/tm/tm.pdf

brands and TAM keywords. The type of correction ranges from acronyms expansion and abbreviations, truncations and grammatical errors which are common in online interactions. The *hunspell* package from R provides a high-performance stemmer, tokenizer and spell-checker in almost any language[5]. It compares each individual word in our data to its built-in US English (en_US) dictionaries - other dictionaries can be added to the working directory. *Hunspell* spell-checker recognizes 9% of misspelled words in the YouCam review data based on a US English dictionary. Though the proportion is not significant to the overall words, they contain informative words that can influence the learning process of ML algorithms in recognizing features.

String distance operations and fuzzy matching are useful tools that are widely used as the basis of automatic document spell check and search queries correction in search engines (Kwartler, 2017). The next chapter will discuss these techniques in greater detail. To speed up the automatic process of correction, we manually reduced the list of corrected words by removing, [1] abbreviated stop words (ie. cuz, bcs, lyk, vry, pls), [2] written expression (ie. haha, ahhh, aww, ew), [3] internet slangs (ie. omg, yay, smh, bam, jk, bff, geez), [4] personal pronouns (ie. me, im, ur , his, hers, ours), [5[other words whose meanings cannot be inferred from the context and only occur once. Once incorrect texts have been replaced, the review text is once again iterated from tokenization.

Once the correction has been performed, we proceed to the process of removing inflected and derived words to their word stem, base of root form of a word that simplifies the term aggregation process for information retrieval (Raghavan, 2008). For instance: “buy, buying, buyers” are all derived from the same word “buy”. There are two methods of stemming and lemmatization to achieve this result. Lemmatization, though proved to be more accurate in text classification (Balakrishnan and Lloyd-Yemoh, 2014), requires part-of-speech (POS) tagging and computationally more extensive. Porter Stemmer[6] is therefore preferred for this analysis and considered sufficient to normalize terms for text summarization purposes.

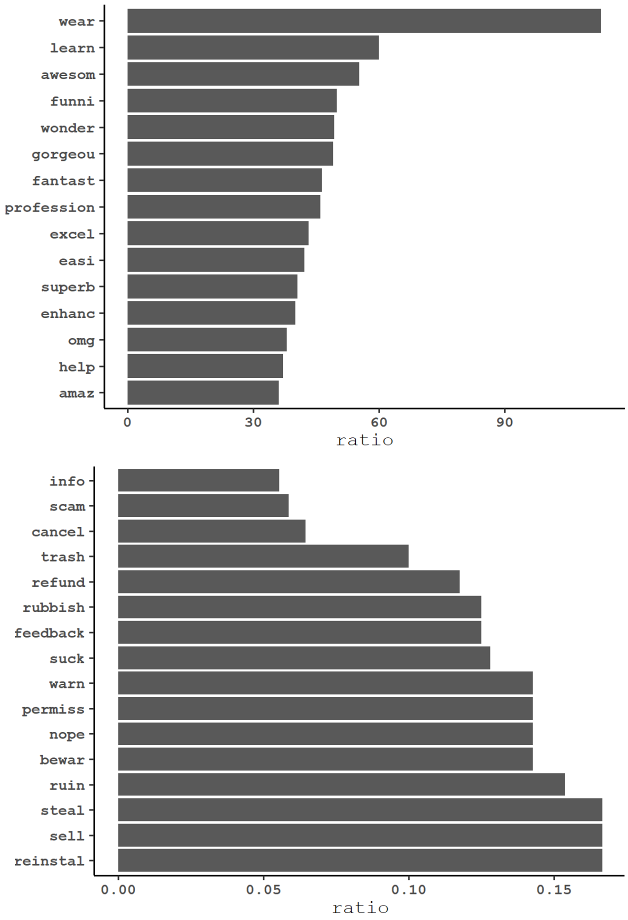
Once the clean review text has been obtained, we compute the relative frequencies of top 15 words used in both positive (≥4 stars) and negative (≤2 stars) users as depicted in the following bar plots. Relative frequency looks at the ratio of term frequencies in a document relative to the

[5] https://cran.r-project.org/web/packages/hunspell/vignettes/intro.html

[6] https://tartarus.org/martin/PorterStemmer/

other document ($TF_{\geq 4stars/}$ $TF_{\leq 2stars}$) which provides a better terms' discrimination within both groups.

Figure 4: *Relative words' frequencies of positive reviews (left) and negative reviews (right)*



(a) (b)

As seen from (a), the most occurring words in the positive reviews relative to negative reviews are dominated by positive adjectives such as "awesome", "funny", "wonderful" and "gorgeous". In contrast, "scam", "trash", "rubbish" and "suck" comprise words that occur mainly in the negative reviews (b). These words may represent opinions or attitudes towards using the app (ATU). Additionally, the words "easy", "enhance", "learn" and "help" in (a) that could signal the positive aspects of PEOU and PU are among the words that occur frequently in positive reviews. In (b) the aspect of BI dominates, as expressed from the words "cancel", "refund", "reinstall" that could imply the unfavorable intention to use the apps. This provides us with an idea of some relevant seed words we can include to extract our TAM features.

3.2 Domain Knowledge

Topic categorization is a subjective task; for different application purposes, different categorization may be required. Due to the aim of this research to build predictive models upon a deductive approach of TAM variables, some domain knowledge will be supplied to guide the topics generated by the algorithm. The semi-supervised seeded-LDA topic modelling[7] allows researchers to predefined topics with keywords, often referred as "seed words" to perform theory-driven analysis of textual data (Jagarlamudi et al., 2012, Watanabe, 2020). The model puts higher weight on these seed words that represent the TAM topics and consequently nudge the model into the desired direction. As a result, the model generates a mixture of user-defined topics and other unsupervised topics and terms that are adjustable via the model's parameters.

The selection of seed words is a two-step process. Firstly, they must help in defining the categories, which is the TAM variables in this case and secondly, they need to contain as little ambiguity as possible. There are two ways in which seed words can be obtain, by a knowledge-based and frequency-based seed words. Watanabe and Zhou (2020) show that a knowledge-based seed words is superior to the latter since they provide operational definitions of the topics and produces a greater external validity to ensure portability across corpora. In selecting the seed words, we also want to incorporate frequency-based seed words we obtained from figure 3 that happen to coincide with our knowledge in defining these TAM topics.

In addition to these words, we consider the definition of TAM variables in table 2. To achieve this, we include [1] the root word of the variables themselves; interact, useful, immerse, vivid and enjoy, [2] words that may provide context to them (ie. color, quality for vividness), and [3] their synonyms based on Oxford English dictionary (ie. fun, enjoy, entertain, excite). Lastly, we also include some words from Rese et al. (2014) that manually coded the TAM variables such as practical and useful for perceived usefulness (PU). Trust and attitudes towards using are excluded from seeded LDA since they will be extracted by sentiment analysis. We tried to distinguish the seed words from one topic and another to prevent overlapping interpretation of the topics.

[7] https://cran.r-project.org/web/packages/seededlda/seededlda.pdf

Table 3: *The selected "seeded words" for semi supervised TAM Topics/Variables*

TAM Topics	Seed Words
Perceived Interactivity	interact, change, adjust, edit, modify, navigate, choose, add
Immersion	realist, real, unreal, unrealistic, accurate, natural, wrong, unnatural, animated
Vividness	vivid, quality, bright, vibrant, detail, color, picture, clarity
Enjoyment	game, enjoy, entertain, excite, play, bore, fun, laugh
Perceived Ease of Use	easy, hard, crash, lag, slow, time, simple, quick, complicated, stuck, confusing
Perceived Usefulness	use, help, learn, enhance, inspire, handy, convenient, inconvenient, practical
Behavioral Intention	recommend, cancel, buy, purchase, delete, download, install, subscribe, uninstall

These seed words above include stemmed words from all parts of speech (POS); nouns, adjectives, verbs that allow the inclusion of both singular, plural nouns, and different conjugations of the verbs. Note that since the primary focus of this research is to find *which (*instead of *how)* TAM's variables are important in determining rating score, we ignore the directionality of the words by including both positive and negative words in each of the TAM's topic, such as "realistic" and "unrealistic" under immersion. We focus on the presence of the seed words in signaling the TAM's topic. Furthermore, Jagarlamudi et al., (2012) emphasizes the importance of including just the "right" number of seed words per topic, as short texts require a great number of seed words to increase their chances in occurring while too many seed words may carry the risk of overfitting. Therefore, I consider 8-10 words as sufficient to guide the LDA but also let the model search for words related to them.

4. Methodology

This section discusses the research methodologies used for text analysis in the systematic order, as well as the reasoning behind selecting them for the purpose of this study. We introduce the first concept of Levenshtein string distance metric for document spell check as a part of data preprocess. For the core analysis we primarily focus on using various Natural Language Processing (NLP) techniques of sentiment analysis, both semi-supervised and unsupervised Latent Dirichlet Allocation (LDA) topic modeling to extract features. Subsequently, these features are used as an input for supervised Machine Learning (ML) models; random forest and multinomial regression to build the predictive models.

4.1 String Matching

In the application of statistical text processing, comparing text strings in terms of their distances is one of the widely used and fundamental tools. String distance quantifies the linguistic similarity of two or more strings on character levels (Kwartler, 2017), for instance the word “book” has the distance of 1 to “cook”. The string-matching function is especially useful for statistical matching, text search, document spell check, text classification and genomics. In the context of this research, we use string distance function for document spell check via approximate string matching. Instead of looking for the location of the match of one string in another string, we use the dictionary approach that finds the location of the closest match of a string in a lookup table.

Loo (2014) identifies three broad categories of string distance metrics, edit-based distances, n-gram based distances and heuristic distances. The type of error in our review data is characterized by typographical error, and more than 80% of these errors deviate from the correct spelling by deletion transposition of two adjacent letter, substitution, or insertion of a single letter (Damerau, 1964, Pollock and Zamora, 1983). Therefore, we select the edit-based distances as the appropriate metric to perform spell check and correction since it counts the minimum number of operations (insertion, deletion, substitution, and transposition) allowed to turn one string into another. (Kashefi et al., 2012, Kwartler, 2017). There are five distinct methods that dictate which of these four substrings operators are allowed, commonly referred as edit distance methods:

Table 4: *Methods of edit strings' distance (Kwartler, 2017)*

Method	Substitution	Deletion	Insertion	Transposition
Hamming	Yes	No	No	No
OSA	Yes	Yes	Yes	Yes, Only Once
Damerau-Levenshtein	Yes	Yes	Yes	Yes
Levenshtein	Yes	Yes	Yes	No
Longest common string	No	Yes	Yes	No

Levenshtein Distance

The selection on either one of this method depends on the judgement of the text miner to infer the required distance calculations from one string to another in an accurate and effective manner (Kwartler, 2017). We select Levenshtein optimal string alignment algorithm that allows the three type of operations from one string to another by substitution, deletion and insertion. Mathematically, Levenshtein distance d_v between string a and b is computed by the weighted number of these operations expressed in the following recursive piecewise functions:

$$d_{lv}(a,b) = \begin{cases} 0\, if\ a = b = \varepsilon \\ min\,\{ \\ d_{lv}\left(a, b_{1:|b|-1}\right) + w_1, \\ d_{lv}(a_{1:|a|-1}, b) + w_2, \\ d_{lv}\left(a_{1:|a|-1}, b_{1:|b|-1}\right) + \left[1 - \delta\left(a_{|a|}, b_{|b|}\right)\right] w_3 \\ \}\ otherwise. \end{cases} \tag{1}$$

In this equation $w_1, w_2, and\ w_3$ represent the nonnegative penalties when $a \neq b$ by substitution, deletion, and insertion. This method is chosen over the rest as misspells in our data are found on either a missing alphabet (i.e., exited for excited), an excess alphabet (i.e., beautifull for beautiful) or a wrong alphabet used (i.e., subscribtion for subscription) that can be corrected via insertion, deletion, and substitution respectively. Though other edit method such as OSA and Damerau-Levenshtein allow a greater number of edit operations, this may increase the possible number of paths between two strings (Loo, 2014), that results in a broader dictionary for string alignment.

4.2 Features Extraction

Features' extraction is the second step of text analysis procedure once the data has been preprocessed. In the marketing academia practice, this process allows researchers to conclude "what is being said?" and "how it's being said?" by exploring the context of the words such as brands, people, product attributes, opinions, and locations (Berger, et. al, 2019). This paper aims to find the most important drivers in predicting the rating score of AR-driven YouCam Makeup app by two folds, a deductive and inductive approach. Firstly, [1] the deductive approach is done by performing a theoretical conceptual analysis built upon TAM and factors suggested from other studies. A conceptual analysis is "a concept is chosen for examination, and the analysis involves, among other things, quantifying and tallying its presence" (Palmquist et al., 1997). Thus, the goal is to verify the validity of TAM variables, perceived interactivity, immersion, vividness and trust in predicting the rating score of the app. Secondly, [2] we use a thematic analysis to complement the theoretical construct [1] by capturing other potential important attributes in the textual data to predict ratings that fully rely on a statistical approach.

In the context of machine learning, the first deductive approach is known as a semi-supervised problem and the second inductive approach is known as an unsupervised problem. By identifying the ML problems, we have narrowed down the list of possible features extraction methods that are appropriate for this research. With the absence of labels on the data, we remove the list of supervised learning techniques on text classification such as conditional random fields, Naïve Bayes deep learning, and Markov models (Berger, et. al, 2019). The next candidates for features extraction fall between the categories of entity (word) extraction and topic extractions.

Entity extractions involve the most basic level of text mining to extract individual entities by words' occurrences, also referred to as a bag of N-grams method. It can be used as an input of a predictive model by generating words associated with a certain class or as dictionaries to extract more complex textual forms such as sentiment or emotion. The N-grams may include unigram, bigram ("credit card") or trigram that capture the number of unordered consecutive words that co-occurred within the same window of document. Rese et al. (2014) in their text mining application of Ikea Place's online reviews adopts this method that has undergone intra-judge reliability to code words associated with each of TAM's constructs; coding 1 for positive association, -1 for negative association and 0 for absence. For instance, to quantify perceived ease of use they coded -1 for words such as "does not work" or "confusing" and 1 for "simple" and "fast". Regardless of its simplicity, the entity (word level) extraction has two major limitations of; [1] dimensionality problem (large number of entities are extracted), [2] interpretation problem due to excessive entities (Berget et al., 2019). Moreover, the interpretation may also be misleading since the scoring relies fully on absolute word frequency; words that occur the most score higher (overemphasized) but may not necessarily contain higher "informational content" to the model as less occurring domain specific words (Brownlee, 2017).

4.2.1 Topic Modelling

To overcome these limitations, topic modelling is another useful feature extraction tool that can generate smaller topics or key themes, composed over a collection of words with a probabilistic-based approach (Kwartler, 2017). When exposed with a vast amount of text, it is particularly useful in automatic text summarization such as document tagging and improving understanding of document context to generate insights (Tirunillai and Tellis, 2014). Furthermore, it can be used in conjunction with other supervised ML models and statistical algorithms for prediction (Geletta et al., 2019). Unlike the traditional clustering, a document can contain multiple topics and terms can belong to multiple documents often referred as 'soft clustering" method, which is more appropriate for this review data; a document (review) can discuss multiple aspects and opinions regarding the apps. Hence, we select this method that can facilitate our concept-based features of extended TAM, by generating unique terms that represent each of them (PU, PEOU, BI, ATU, trust, enjoyment, vividness, and immersion) and by summarizing topic probabilities for each document (review). The two most used topic models are derived from the predominant approach of support-vector-machine Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA) and Poisson Factorization (PF) (Blei et al., 2003).

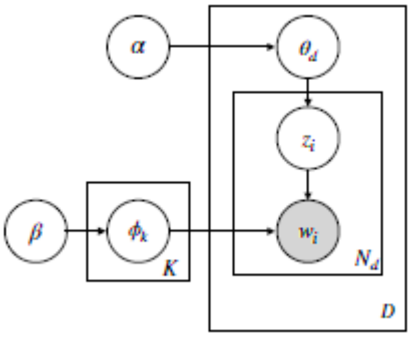
4.2.1.1 Latent Dirichlet Allocation (LDA)

As one of the foundational approaches in text mining, Latent Semantic Analysis (LSA) uses document-term matrix (DTM) as the data input with rows correspond to documents, columns represent unique words in the vocabulary, and values represent the words' occurrence frequency. (Berger et al., 2019, Deerwester et al., 1990). These values can be in the form of Term Frequency-Inverse Document Frequency (TF-IDF) that rescale the frequency of words' occurrences by penalizing frequent words that are also frequent across all documents, thus it emphasizes terms' importance rather than frequency based on their informational content (Kwartler, 2017). Hence, it overcomes the overrepresentation problem that arises from using absolute term frequency in the bag of words approach. This matrix is then decomposed into two lower dimensional matrices through Singular Value-Decomposition, resulting in topic-term probabilities (ϕ matrix) and document-topic probabilities (θ matrix).

Derived from this principle, Latent Dirichlet Allocation (LDA) is a more refined and popular generative probabilistic approach of LSA that can be generalized to unseen documents (Blei et al., 2003). "Latent" refers to the model's ability to uncover hidden topics and hidden terms' representation on each topic (outside the knowledge of the text miner) and it places Dirichlet distribution to model seemingly random allocations of the words (Kwartler, 2017). LDA with the combination of supervised ML using the document-topic probabilities as an input, has demonstrated an improvement in predictive accuracy within various applications; classification of phishing messages (Ramanathan and Wechsler, 2013, prediction of clinical trial termination (Geletta et al., 2019), drug reactions (Xiao and Chang, 2017). Moreover, LDA's popularity results in various model's extension arise from many contributors, offering a great flexibility on its application. One of them includes the semi-supervised version of LDA that allows incorporation of domain knowledge (theoretical construct) which is highly applicable to generate TAM variables. Altogether, the promising predictive accuracy, availability of a semi supervised LDA, and the appropriate level of complexity and interpretability make LDA as the most suitable candidate for this study.

LDA's application holds the following assumption: the number of topics is known a-priori and it ignores order of words and grammatical structure. As mentioned, another fundamental assumption of LDA is that each topic is composed over a collection of words and that each document belongs to multiple topics. To start with, a few worth mentioning annotations for the LDA applications in this research include, [1] a document (D) refers to a sequence of N words that compiles a review, [2] A term (w) refers to a word as a single item from a vocabulary, and [3] topic (K) refers to the summarization of words that define a feature. We let ϕ to denote the matrix of topic-term proportions and θ to denote the matrix of topic-term proportions with rows summing up to 1. This process of topics and words assignment to each topic in a corpus is graphically represented as follows:

Figure 5a: *Plate notation of LDA's model parameters*



Where the outer box (D) represents the number of documents and the inner box represents the iterative process of topic and word assignment within a document. The Dirichlet priors of α and β are parameters at a corpus level involved in the sampling process. Dir(α) dictates the density of per-document topic distribution ϕ_k; a high value of α implies a high mixture of topics within a document. Dir(β) dictates the per-topic word distribution θ_d, similarly; a high value of β implies a high mixture of words. Whereas z_i is the assigned topic for the *n*-th word in document *D*. The generative process of LDA on each document *w* in a corpus *D* involves three levels that discriminates it from a two-level classical clustering model. It starts by a random assignment of weights to both matrices θ and ϕ. Subsequently, we randomly assign a topic across the distribution of topics to a document based on their assigned weights. Lastly, we assign a word from the distribution of words to the selected topic at random. These processes are repeated for the entire document until it reaches converge and can be illustrated by the following steps (Blei et al., 2003):

1. For each document d ∈ {1, ..., D} choose $\theta_d \sim$ Dir(α)
2. For each topic k ∈ {1, ..., K} choose $\phi_k \sim$ Dir(β)
3. For each of the N tokens w_i , i ∈ {1, ..., N_d} in each document d ∈ {1, ..., D}:
    a. Select a topic $z_i \sim$ Multinomial (θ_d).
    b. Select a word $w_i \sim$ Multinomial (ϕ_k).

In this data generating process, we aim to estimate the weight that maximizes the likelihood of our data given both matrices. There exist two main algorithms that help in identifying the correct weight; Expectation-Maximization and collapse Gibbs sampling (Blei et al., 2013; Griffith and Steyvers, 2004). In this research, we use the Markov chain Monte Carlo; Gibbs Sampler algorithm that is relatively simple and fast to work on large corpora (Steyvers and Griffiths, 2007). It is indirectly approximate the probability distribution of a single word allocation (*w*) as a part of document (*D*) to a topic (*K*), conditional to the rest of the topic assignments to estimate its latent variables; ϕ_t (topic-term) and θ_d (document-topic) via several sampling iterations. Mathematically, we can express this in a conditional posterior distribution equation of z_i as follows (Griffith and Steyvers, 2004):

$$P\left(z_i = j | z_{-i}, w_i, d_{i,} \cdot\right) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha} \tag{2}$$

Where,

z_{-i} = Assignment of all z_k such that $k \neq I$

$n_{-i,j}^{(w_i)}$ = Number of words assigned to topic *j*

$n_{-i,j}^{(\cdot)}$ = Total number of words assigned to topic *j*

$n_{-i,j}^{(d_i)}$ = Number of words from document d_i assigned to topic j

$n_{-i,\cdot}^{(d_i)}$ = Total number of words in document d_i

The above equation consists of two parts, the first part represents the presence of topic in a document and the second part represents how fitting a word is in a topic. The Dirichlet priors of α and β act as a smoothing parameter when $n_{-i,j}^{(w_i)}$ and $n_{-i,j}^{(w_i)}$ are equal to zero, which keeps the likelihood of the word's presence in a topic. The values of z_i are initialized from 1 to T, then The Markov chain ran for several iterations until the target distribution is reached and the values of z_i are captured.

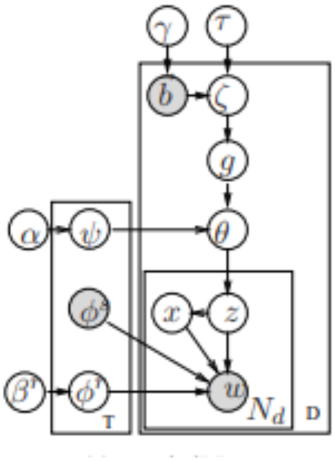
4.2.1.2 Semi-supervised LDA

LDA is by principal considered as an unsupervised NLP technique that fully relies on statistical inference to generate topics. This process has a tendency to results in many non-specific terms or "junk" topics, governed by the phenomenon of "higher-order co-occurrence" (Heinrich, 2009). A degree of supervision can address this issue and strengthen topics' interpretation to generate richer managerial insights (Tirunillai and Tellis, 2014). LDA allows extension to a semi-supervised form by incorporating domain knowledge to influence the topic generated by the model. For this reason, LDA is chosen as the baseline model to generate latent dimensions of our TAM variables. In this case, the topics are defined by the mixture of two Multinomial distributions: regular topic $\phi_k^r$and seeded topic ϕ_k^s. To achieve this, seed words from *Table 3* are added to the model's parameter which can be controlled by an additional hyperparameter of *seed confidence* π_k. Seed confidence refers to the strength of domain knowledge we would like to incorporate in the model, by adding an extra boost to these terms as part of the topic. Furthermore, we control for the list of set seeds (binary vector of $\vec{b}$) that are allowed for each of the topic-word distributions ϕ_k and S group-topic distributions ψ_s In contrast with unsupervised LDA, for words w_j in the predefined set W_j ($w_j \in W_j$), the value of z_i is known and remains constant throughout the Gibbs sampling iterations. Consequently, this whole generative process extends to the following steps (Jagarlamudi et al., 2012);

1. For each document d ∈ {1, ..., D}
    a. Choose $\theta_d \sim \mathrm{Dir}(\psi_g)$
    b. Choose a binary vector $\vec{b}$ of length S
    c. Choose a document-group distribution $\varsigma^d \sim \mathrm{Dir}(\tau\vec{b})$
    d. Choose a group variable $g \sim \mathrm{Multinomial}(\varsigma^d)$
2. For each topic k ∈ {1, ..., K}
    a. Choose regular topic $\phi_k^r \sim \mathrm{Dir}(\beta_r)$
    b. Choose regular topic $\phi_k^s \sim \mathrm{Dir}(\beta_s)$
    c. Choose $\pi_k \sim \mathrm{Beta}\ (1,1)$
3. For each seed set of s ∈ {1, ..., S} Choose $\psi_s \sim \mathrm{Dir}(\alpha)$
4. For each of the N tokens w_i, i ∈ {1, ..., N_d} in each document d ∈ {1, ..., D}:
    a. Select a topic $z_i \sim \mathrm{Multinomial}\ (\theta_d)$.
    b. Select an indicator $x_i \sim \mathrm{Bern}\ (\pi_{zi})$ if
        i. $x_i = 0$ select a word $w_i \sim \mathrm{Multinomial}\ (\phi_k^r)$.
        ii. $x_i = 1$ select a word $w_i \sim \mathrm{Multinomial}\ (\phi_k^s)$.

And graphically represented,

Figure 5b: *Plate notation of seededLDA's model parameters*



4.2.2 Sentiment Analysis

Sentiment analysis is another useful tool to quantify the emotional intent such as attitudes, thoughts or judgements that can influence a certain behavior (Kwartler, 2017). The sentiment is made up of a couple elements: "an option [1] of a person, [2] about a target (or aspects of the target), [3] that has a certain valence (+/-/0) or emotion, and [4] strength. There are numerous frameworks for sentiment analysis that are derived from two prominent approaches of lexical-based and machine learning approach (Dhaoui et al., 2017). Lexical-based approach is a widely used sentiment classification due to its simplicity; it uses the bag-of-words approach that solely relies on the occurrent of a word and ignores their order. Machine leaning approach, on the other hand uses a more complex human-coded trained data that tends to produce a more accurate sentiment classification (Hartmann et al., 2018; Borah and Tellis, 2016). The decision to select one of these approaches should carefully consider the trade-off between empirical and theoretical fit.

In this research, sentiment analysis acts as the secondary feature extraction method to extract only trust and ATU which are measures of users' emotional intent. We consider the lexical-based approach sufficient to generate these features as ML approach requires intensive resources to manually label our data. Furthermore, in Dhaoui et al. (2017) shows that the two approaches have similar performance in the context of social media conversations. We complement our lexical-based approach with contextual polarity that consider valence shifter surrounding the focal word such as *"not"* or *"very"* for the word *"good"* that change the semantic meaning of the word itself. Consequently, moderate the limitation of the bag-of-words approach.

4.3 Prediction Models

The features generations methods result in two data frames, [1] DF-sLDA that consists of topic-probabilities on each of TAM, sentiment scores, length of the review and review time attached to each text review, and [2] DF-LDA that consists of topic-probabilities on each of the generative topics, sentiment analysis, length of the review and review time on each text review**.** To achieve the main objectives of this research, we use the two datasets as inputs to our supervised models to solve classification problem of 5 rating classes. The proposed classification model needs to consider, [1] the nature our data which is highly skewed to the rating point of 5, [2} the existence of outliers, [3] highly correlated features, [4] the number of features and [5] the appropriate balance between computational complexity and prediction accuracy.

This list of considerations has led to the selection of ensemble methods. This method combines multiple nonlinear weak classifiers that often lead to better results than a single complex classifier such SVM or ANN while retaining runtime computation (Guestrin and Chen, 2016). Random Forest in particular, is one of the most widely used bootstrapped aggregation (bagging) for ensemble method that combines several decision trees, resulting in a relatively stable (prone to outliers and overfitting) and accurate model. Furthermore, it minimizes the distortion of correlated features by randomly search for important subset of features in each node while growing the trees. This feature is therefore, particularly useful when correlations among features are expected which is applicable for our LDA features that contains composite nature. The row of LDA matrix sums up to 1, which means that an increase in one of the topic probabilities decreases other(s).

The theoretical groundwork of RF has fulfilled the criteria of our preferred classification model, with the exception of [1] the problem of imbalance class, which may cause prediction bias towards the majority class. Though RF performs better than linear models in overcoming this issue by assigning weight to the previously misclassified data in the next iteration, the problem persists to a large extent due to the large variability between the training and test set for minority classes sampling (Blagus and Lusa, 2010). Several literatures have shown that the addition of balancing techniques such as under sampling, oversampling and synthetic oversampling approach (SMOTE) can significantly address this issue and improve the base classifiers. Though it is beyond the scope of this research, this could be integrated in the future research.

4.4 Evaluation Metric

We are interested in evaluating the performance of our feature generation models in terms of their predictive performance. This is done in two ways; firstly, by measuring the ability of our topic generation model to generalize unseen data and its performance as input to our rating predictive models. Evaluating topic models is challenging due to its unsupervised nature, especially in NLP in which the interpretation tends to be vague and subjective. Yet it remains as an equally important task to compare different models with different number of topics and hyperparameters. Perplexity is a widely used metric for language model evaluation that can objectively use to compare different models. At the predictive level, the external evaluation of our features generation models is comparable to evaluating clusters. The documents (soft) clusters resulted from the topics can be used to predict our rating class. Clusters evaluation metrics vary from the adaptation of classification metric or information theory. Since the focus of this research is to predict classes, we use the classification metrics as the basis to our (soft) clustering metric. Thus, the measure of perplexity is used at topic generation level and classification metrics are used at the predictive level.

4.4.1 Perplexity

As previously mentioned, LDA assumes that the number of topics is known a-priori. Though there is currently no formal way in determining how many topics should be extracted using LDA, the measure of perplexity can be used to find the optimal number of latent topics. It is also a common evaluation metric to compare different probabilistic model. Perplexity measures the efficiency of the model performance on the held-out test data, and it is calculated as the geometric mean per-word likelihood (Newman et al., 2010). However, some studies show that perplexity may not yield to human interpretable topic (Chang et al, 2009). Their study results in the counterintuitive findings in which humans preferred model with worse measure of perplexity. Thus, though we cannot fully rely on the topics generated by perplexity, it is nonetheless another useful metric in topic modelling toolset and will be used to compare the optimal model's parameters.

4.4.2 Classification Metrics

Selecting the right metrics for classification problems are a challenging task due to the large number of available metrics. In addition to the standard classification errors, we need to consider the characteristic of our data and apply the appropriate metrics that lead into accurate interpretation. In this case, our data contains a highly imbalanced class of 5 stars rating. To obtain an initial impression of overall performances comparison across models, we will compute the metric of *accuracy* that calculates the correctly predicted instances over the total predicted instances, *recall* (sensitivity) and *precision* (specificity). These standard accuracy metrics, though works as a good indicator for most cases assume balanced class distribution. The distribution of our dependent variable of rating imposes the problem of imbalanced class that is highly dominated by rating of 5, causing the level of accuracy that is misleading towards this class. Literatures suggest that metrics such as *AUC, F1* and *Cohen's Kappa* are better measurements for this problem (Jeni et al., 2013). *Cohen's Kappa*, in particular, is a measurement of chance and captures imbalanced class problems into account, resulting in a more reliable measurement for our datasets. Jeni et al. (2013) shows that in the application of imbalanced class, *Kappa* in comparison to the other metrics is more tolerant to the prior distribution of the classes. This led us to the usage of following classification metrics to evaluate the performance of our RF and multinomial logit models:

$$Accuracy = \frac{Correctly\ Predicted\ Instances\ (TP + TN)}{Total\ Predicted\ Instances\ (TP + FP + TN + FN)}\ x\ 100 \tag{3}$$

Measures the overall accuracy of correctly classified instances over total predicted instances,

$$Sensitivity = \frac{TP}{(TP + FN)}\ x\ 100 \tag{4}$$

$$Sprecificity = \frac{TN}{(TN + FP)}\ x\ 100 \tag{5}$$

(4) measures the ability to correctly identified true positive instances and (5) measures is the ability to correctly identified true negative instances.

$$Cohen's\ Kappa\ = \frac{P_{observed} - P_{chance}}{1 - P_{chance}} \tag{6}$$

Kappa measures the inter-rater reliability of the degree of agreement between two or more raters (< 0.2 poor agreement, 0.21 – 0.40 fair agreement, 0.41 – 0,60 moderate agreement, 0.61 – 0.80 good agreement and 0.81 – 1.00 very good agreement) (Parraga-Alave et al., 2021).

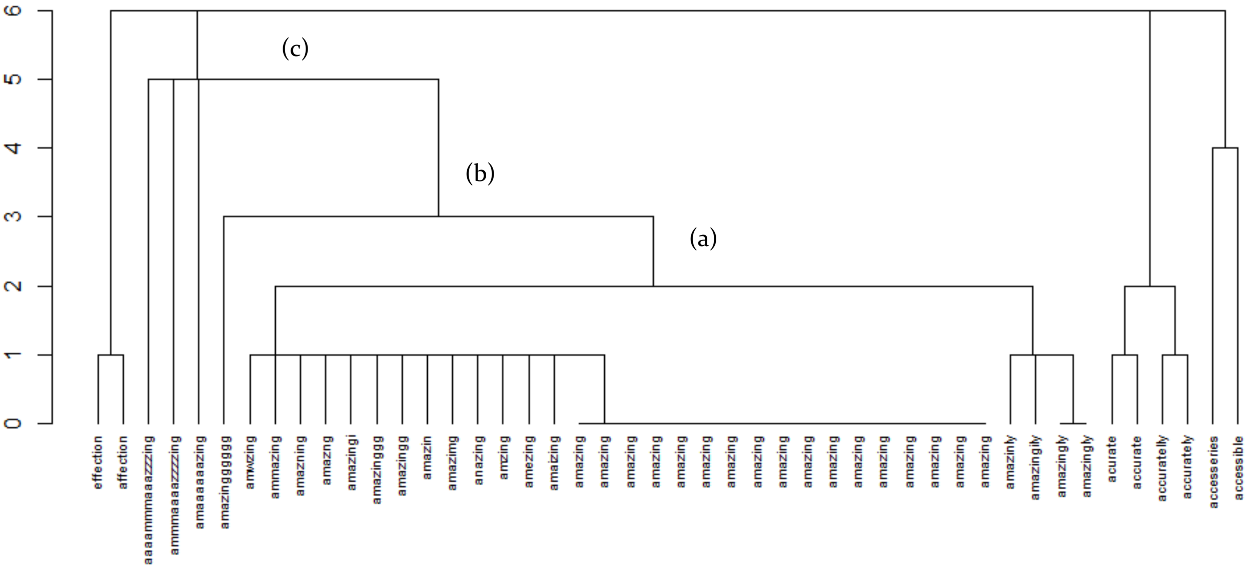
5. Implementation

This section guides the implementation of our selected methods; Lavensthein optimal string alignment to correct misspells, features extractions using LDA and sentiment analysis to the preprocessed YouCam Makeup review data.

5.1 Lavensthein Optimal String Alignment

We retrieve a custom US English Dictionary from Qdap Dictionaries[8] in package *qdap* that contains a list of 20137 US English words. We merge this vector with the list of correct tokenized words (since some words are missing in this dictionary) and words that do not exist in the dictionary but have high occurrences and may provide context such as “app” and “makeup”. Subsequently, we create a vector of 820 misspelled words that we are interested to recover (TAM seed words, valence shifters, and adjectives) and a correction function. This correction function is built upon the *adist*[9] function that computes the approximate Levensthein string distance between the two vectors and find those that give the minimal Levenstein edit distance (the weighted number of all allowed operations). In other words, the function proposes a correction to our custom dictionary that are closest to each of the misspelled word in Levenstein distance; the dendrograms below depict the words “amazing” and “accurate” as an example to illustrate this process.

Figure 6: *Words with the closest Levenstein distance to “amazing” and accurate”*



[8] https://cran.r-project.org/web/packages/qdapDictionaries/qdapDictionaries.pdf

[9] https://cran.r-project.org/web/packages/stringdist/stringdist.pdf

Cluster (a), shows the misspelled words of “amazing” with the distance of 1, meaning only 1 operation needed to convert these words to “amazing” (i.e., deletion of *m* for “ammazing”, insertion of *g* for “amazin”, and substitution of *a* to *e* for “amazing”). The subsequent clusters (b) and (c) consist of words with further distance but still within the boundary of Levenstein operations. Since we cannot fully rely on all the proposed correction generated by the algorithm, we briefly checked at the corrections and performed a manual correction when necessary. The correction has reduced the proportion of the misspelled words to 5%. The final step is to replace all the misspelled words with their correction in the data frame which will be utilized for the next step of the analysis.

5.2 LDA Topic Models

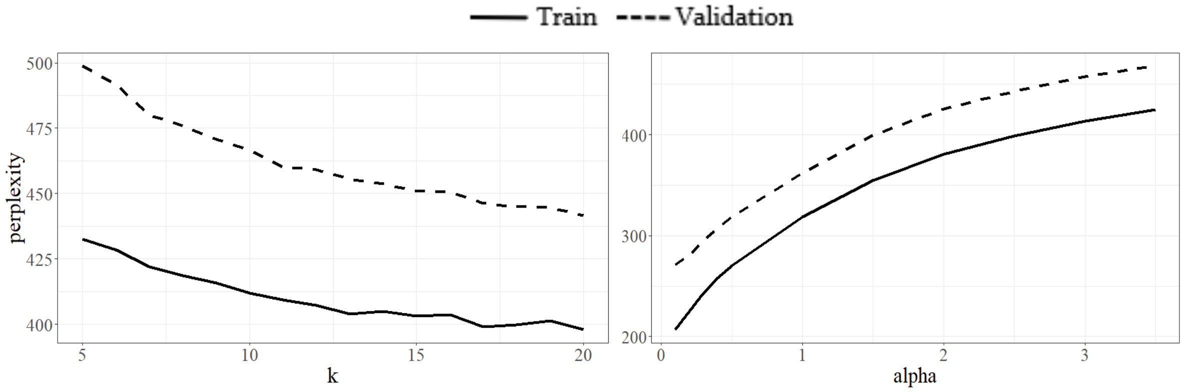
To initiate the LDA topic models, we converted the preprocessed review text data into document-term matrix using the *quantenda*[10] package and remove words with less than 6 words to limit the sparsity of our data, resulting in 11170 observations left to be analyzed. The problem of sparsity occurs when dealing with short documents due to the difficulty in identifying ambiguous words within a limited context. Consequently, the model tends to assign only one single topic to the short document that affect its flexibility and risk of overfitting (Tan et al., 2013).

We proceed with LDA topic models to generate TAM features as our baseline model using *seededLDA*[11] and unsupervised features using *topicmodels* respectively from R. Both models fitting requires the number of topics (K) and a few hyperparameters of *alpha*, *beta* and *seed confidence* (only for semi-supervised LDA) to be initialized. *Seed confidence* is a constant of $0 < \pi < 1$ that accounts for the strength of domain knowledge, calculated as the pseudo count given to the seed words as a proportion of the total number of words. Griffith and Steyvers (2004) suggest a value of 50/K for α and 0.1 for β, not for an arbitrary reasoning but to account for the number of topics. We adopt their suggestion on setting the βs = 0.1 for both models to achieve fine-grained topics that address specific areas. But since we don’t know the number of topics a-priori for the unsupervised LDA model, the optimal hyperparameters can also be obtained in a data driven way like any other machine learning models by finding those that maximize the approximate log-likelihood of the models.

[10] *https://cran.r-project.org/web/packages/quanteda/index.html*
[11] *https://cran.r-project.org/web/packages/seededlda/seededlda.pdf*

We start by initializing the seededLDA model. Quoted from the author, “Seeded-LDA allows users to pre-define topics with keywords to perform theory-driven analysis of textual data”. For our seeded LDA model we predefined K = 8, 7 number of TAM topics (except for ATU and trust) and 1 topic of *others* for topics that don’t belong to any of the 7 TAM categories as a baseline to control occurrence of words. We set the *residual* = *TRUE* to let the model finds related terms in addition to our seed words and compose the terms in *others*. Since our goal is to compare various features of YouCam apps in terms of their predictive power to rating, we want to limit the overlap of TAM topics and put higher emphasize on the dominating topics within a review by controlling for the document-topic density (α) to 0 and account for equal strength of our domain knowledge and the algorithm by setting the π of 0.5.

Figure 7: *Words with the closest Levenstein distance to “amazing” and accurate”*



(a) (b)

In contrast with seeded LDA, we rely on the statistical approach for our unsupervised LDA to find K and α by performing grid search on range of values with the lowest perplexity scores on the test data. To achieve this, we split the document-term matrix data to 80:20 train/validity set and create a perplexity function on LDA for the range of the first 20 topics. Though this may appear somewhat arbitrary, the decision to stop at 20 topics was primarily motivated by, [1] the computational cost and [2] the perceived lack of interpretable topics for more than 20 topics, since there will be many overlapping terms across topics. The perplexity converges after it reaches 15 topics, resulting in the number of 20 topics with the lowest perplexity score. Similarly, we repeat this operation to find the optimal α for a range of value of (0.1 – 50/K) on the validation set for K = 20. On the contrary with what Griffith and Steyvers (2004) have suggested, we found that lower bound of α appears to give the lowest perplexity score on the held-out test set and therefore we select 0.1 for the final unsupervised LDA model.

Generally, Gibbs sampling theoretically ensure convergence according to Darling (2011). However, he also pointed out in practice, convergence diagnosis is a real challenge in Gibbs sampler approximate inference since it is difficult to determine the number of required iterations. Thus, to ensure stable results, we ran both models with a burn-in of the sufficiently high number of 2000 Gibbs sampling iterations.

5.3 Sentiment Analysis

The operationalization of **trust** and **attitudes towards using (ATU)** are based on sentiment analysis of users of the YouCam Makeup Apps. Trus**t** is quantified based on the NRC Lexicon dictionary from package *syuhzet* in R that consists of Liu words list with approximately 6800 words signaling the feelings of trust, anger, surprise, sadness, positive, negative, joy, fear, disgust, anticipation, and anger[12]. The value consists of the count of words associated with the feeling of Trust according to this dictionary (i.e., consistent, insure, and integrity). For dataset 1 (TAM), the rest of the emotions' lexicons are omitted.

ATU on the other hand, is obtained by computing contextual polarity scores from package *qdap* in R[13]. Polarity score is a more accurate representation of capturing sentiment than just accounting for positive and negative words from the NRC lexicon. They capture valence shifters, words that change the semantic orientation of the adjective (not good < good, very good > good). We chose to run polarity scores that account for valence shifters of 2 words before and 2 words after the focal (adjective) word. By skimming through the review text, this window is considered appropriate in capturing negation, de-amplifier and amplifier that surround the adjectives and change their meanings. For example, the sentences of *"I really really like this app"* contains 2 amplifier words preceding the positive word of "like" and *"This app is not very fast"* contains a negation word and an amplified word preceding "fast". The polarity score, package qdap in R adjusts for negations, amplifies and length of the sentence by the following computation:

1. Find word that appears in dictionary (=polarity word)
2. Consider 2 preceding &2 following words
3. Positive polarity word = +1 (negative -1)
4. Flip sign for negation word

[12] *https://search.r-project.org/CRAN/refmans/textdata/html/lexicon_nrc.html*
[13] *https://cran.r-project.org/web/packages/qdap/index.html*

5. Add/subtract 0.8 for every amplifier
6. Sum over all polarity words
7. Divide by $\sqrt{no. of\ words}$

5.4 Review Time

Since we would like to investigate the effect of pandemic to AR VTO usage, we operationalize the time of the review as a supporting variable in the POSIXct format (hour: minute day: month: year), split into two categories of pre corona and post corona. The benchmark that separates these categories is 15th of March 2020; around the time when US declared its first Corona death and announced its firs lockdown of the country. We expect that post-corona time frame will boost the aspects of PEOU and PU, as the application helps in facilitating makeup testers as an alternative for users that cannot or choose not to go to the physical stores. Moreover, a survey from Digiday (2021) shows that YouCam Makeup apps experienced an increase of 32% of downloads since the start of the pandemic. Based on this, we expect to see interactions between corona time with BI and PU which will be tested via our prediction model.

5.5 Prediction Models

Once we have wrangled the data frames, we used them as inputs for our RF prediction models. The features generation process has resulted in 4 different data frames that we are interested to measure; model A1[14] comprises of TAM variables via seededLDA, polarity score and corona time, model A2[15] comprises of TAM via seededLDA, NRC Emotions (anger, trust, fear, anticipation, trust, surprise, sadness, joy, disgust, positive and negative feelings), corona time and length of reviews, model B1[16] comprises of LDA topics via unsupervised LDA, trust, polarity and corona time, and lastly, model B2[17] comprises of LDA topics, NRC emotions, corona time and length or reviews. The reasoning behind the selection of these data frames, is that we are primarily interested in the performance comparison between TAM concept-based features (seededLDA, trust and polarity) and thematic features (LDA, NRC and other relevant text features).

The first step is to split the dataset into 80:20 training and test set. We then perform the necessary adjustment to the type of variables, ratings and corona time as factors with levels. We initialized all RF models with R package *randomforest* default hyperparameters (ntrees =500,

[14] Appendix A3 table 9
[15] Appendix A3 table 10
[16] Appendix A3 table 11
[17] Appendix A3 table12

mtry = $\sqrt{p}$, where p is the number of variables available to split in each node). After running the default models, we plot the out of bag error at each trees' iteration (appendix) and found that for all models the errors reach convergence after 100 number of trees. Thus, we select 100 number of trees and mtry = $\sqrt{p}$ as our final models. The last step is to evaluate the performances of these RF models by obtaining prediction to the test set and obtain confusion matrices of correctly predicted classes.

6. Results

This section provides the evaluation of our features' generation methods and classification performances of the selected models. We look at the resulting topics and terms produced from both model A seededLDA and model unsupervised LDA and the sentiment analysis for ATU and trust. We finally compare the performances of these models in terms of their quantitative performance of accuracy measures in our classification models and qualitative performance of topic coherency. The final interpretation of our analysis will be based on the model with the best performances in these regards.

6.1 Qualitative Evaluation

The first step to evaluate our topic models is by looking at the resulting topics and their associated terms. Table 5a and 5b provides the top 10 terms and topics resulted from model seededLDA and model unsupervised LDA respectively. The (*) in table 4a annotates the terms and topics that are not pre-determined and solely produced by the LDA algorithm. The order of the terms (top to bottom) are arranged based on those with the strongest association to the topic.

Table 5a: *Terms and Topics generated from Model A: seededLDA*

PI	Immersion	Vividness	Enjoyment
Change	Natural	Color	Enjoy
Choose	Realistic	Quality	Excite
Adjust	Wrong	Detail	Entertain
Navigate	Accurate	Bright	Laugh
Modify	Unnatural	Clarity	Awesome
Interact	Unrealistic	Vibrant	Excellent*
Option*	Unreal	Vivid	Super*
Choice*	Photo*	Picture	Application*
Hairstyle*	Editor*	Perfect*	Friend*
Feature*	Effect*	Style*	Experience*
PEOU	**PU**	**BI**	**Others***
Simple	Learn	Download	Picture
Crash	Inspire	Recommend	People
Lagging	Practice	Install	Pretty
Quick	Handy	Uninstall	Friend
Stuck	Feature*	Delete	Stuff
Beauty*	Update*	Purchase	Awesome
Camera*	Version*	Cancel	Beauty
Application*	Premium*	Phone*	Creative
Picture*	Eyebrow*	Money*	Person
Editor*	Remove*	Subscription*	World

The terms produced by LDA in our model A yields more coherency in some topics and less in others. For instances, *option*, *choice, hairstyle* and *feature* are related to the nouns associated to PI, whereas *excellent, super* and *application* are not as coherence in defining Enjoyment. These words are more associated with the broader context of positive sentiment. Similarly, the words *camera*, *application* and *picture* in PU provide relatively vague associations to PEOU compared to beauty that can be stemmed from the word *beautify/beautifies.* The same case applies for the words *update*, *version*, and *premium* in PU. The topic of others appears to contain the social and entertainment aspect of YouCam apps based on the terms it produces.

Table 5b: *Terms and Topics generated from Model B: unsupervised LDA*

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Beauty	Update	Love	Love	Color
Make	Change	Friend	Amazing	Lipstick
Love	Version	Girl	Star	Lip
Amazing	Eyebrow	Amazing	Awesome	Add
Feel	Feature	Real	Pretty	Love
Real	Option	Perfect	Download	Adjust
Pretty	Fix	Awesome	Rate	Nice
Pic	Star	Fun	Perfect	Hair
Picture	Total	OMG	Wow	Eyelash
Camera	Hate	Cool	Install	Perfect

Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
Fun	Photo	Hair	Free	Love
Easy	Edit	Color	Feature	Real
Lot	Nice	Change	Pay	Skin
Love	Editor	Style	Premium	Help
Play	Picture	Love	Favorite	Buy
Option	Quality	Hairstyle	Option	Fun
Enjoy	Video	Option	Version	Product
Time	Picture	Amazing	Money	Life
Choice	Perfect	Fun	Purchase	Wear
Choose	Image	Cool	Paid	Idea

Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
Love	Nice	Photo	Love	Game
Edit	Beauty	Love	Amazing	Love
Photo	Camera	Edit	Recommend	Fun
Picture	Love	Feature	Easy	Play
Pic	Picture	Remove	Excellent	Cool
Help	Camera	User	Application	Amazing
Take	Pretty	Option	Highlight	Download
Perfect	Perfect	Filter	Result	Bore
Make	Enjoy	Effect	Awesome	Enjoy
Beauty	Application	Friendly	Realist	Lot

Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
Free	Time	People	Love	Love
Day	Download	Post	Easy	Amazing
Money	Phone	Download	Awesome	Absolute
Time	Take	Bad	Nice	Awesome
Subscription	Load	Picture	Cool	Picture
Charge	Fix	Start	Filter	Beauty
Cancel	Install	YouCam	Super	Easy
Trial	Slow	Pretty	Effect	Favorite
Pay	Uninstall	Delete	Fun	Fantasy
Refund	Freeze	Stuff	Add	Perfect

Model B unsupervised LDA from table 4b, in comparison, yields a more ambiguous terms and topics. This can especially be seen in the most probable terms defining topic 3, 4, 11 12 1,4 19, and 20 that contain a lot of overlapping terms of positive adjectives such as *love, amazing, awesome, perfect*, and *pretty* though in topic 11 and 12 we can see words such as *edit, camera, photo*, and *picture* that are more associated to the feature of the YouCam apps. Within the topics, words also appear less coherence in respect to TAM. For instance, topic 4 contains partly positive ATU and BI expressed by the words *download*, *rate* and *install*. Similarly, topic 6 is composed of words defining Enjoyment and PU. Nevertheless, some topics yield a more interpretable topic such as 5 (PI and products), 16 (monetization) and 17 (PEOU). The most recurring posterior distribution of topics in model B[18] varies, with slight emphasize in topic-probabilities of 04-0.6. This indicates that some topics dominate in most reviews, and some are equally distributed.

6.2 Quantitate Evaluation

The matrices of topic probabilities resulted from both models, polarity score and review time are now used as inputs for our predictive models. As mentioned, we use RF models to predict rating classes. In addition to the qualitative aspect of topic coherency and interpretation of both topic models in 6.1, we are now interested in the predictive power of model A and model B. Both models are treated separately due the composite nature on the topic probability values that make up to 1. Additionally, we ran a multinomial logit regression (model C) to test our expectation regarding the interactions between corona time with BI and PU, as well as PU and PEOU.

[18] Refer to appendix

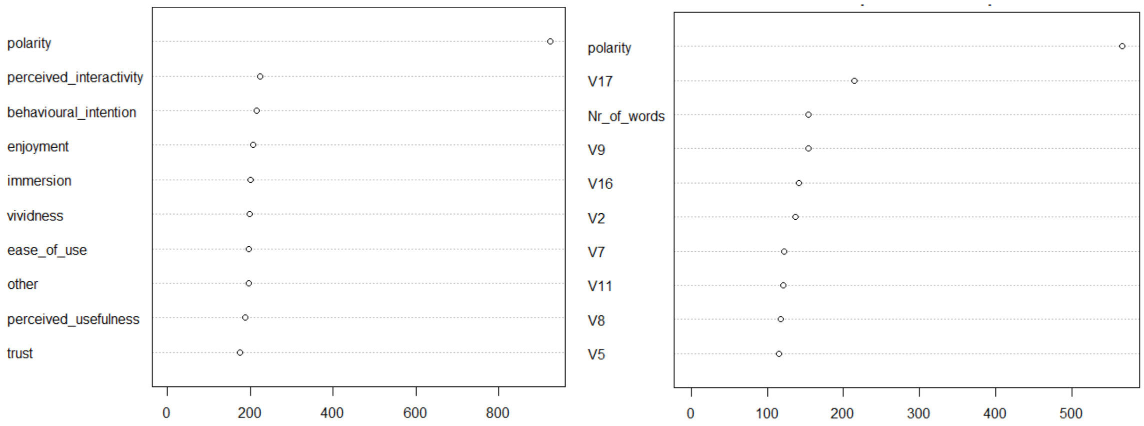
Table 6: *Models Comparison Based on RF (ntrees =100)*

Model	Method Used to Generate Topics	Model Used to Run Prediction	Accuracy	Kappa
A1. TAM + Trust + Polarity + Corona Time	Semi-supervised LDA	Random Forest	0.74	0.26
A2. TAM + NRC Emotions + polarity + Corona Time + No. of words	Semi-supervised LDA	Random Forest	0.74	0.27
B1. LDA + Trust + Polarity + Corona Time	Unsupervised LDA	Random Forest	0.75	0.31
B2. LDA + NRC Emotions + polarity + Corona Time + No. of words	Unsupervised LDA	Random Forest	**0.78**	**0.32**
C*. Model A1 + Corona*BI + Corona*PU + PU*PEOU	Semi-supervised LDA	Multinomial Logit Regression with Interactions	0.75	0.25

As we see from above the accuracy of our 5 models do not differ a lot, lie within the acceptable range of ± 75%. Though the overall accuracy is high, the Cohen's Kappa score performs significantly poor. This is because Kappa takes imbalance class distribution into consideration; the smaller the difference between the distribution of predicted and actual classes, the bigger the Kappa score is providing the more realistic view on performance indicator of imbalanced class. This can also be seen in table 10[19], the rating score of 5 scores the highest sensitivity among all classes, which implies that the model predicts the greatest number of correct 5 rating star reviews than the others. Altogether this gives us model B2 as the best performing model in which will be used to obtain our final interpretation. For comparison purposes in terms of interpretability between our concept based and thematic topic models, we retrieve the 10 most important variables of model A1 (TAM via seededLDA) and the winner model B2 (unsupervised LDA)

[19] Refer to appendix A3

Figure 8: *Top 10 important variables model A1 (left) and model B2 (right)*



The variable importance plots above represent the important features of the YouCam Makeup Apps that are important in determining rating score. Both models produce the highest importance of polarity score (ATU) as the most important predictor, intuitively it captures positive sentiment in the review that leads to higher rating and vice versa. Additionally, polarity (ATU) is proven to be a better measure than the unigram NRC dictionary approach since they consider valence shifter that can change or stress the lexical context of the review.

Our concept TAM based model A1 generate the next best important variables of PI, BI, enjoyment and vividness, though the relative importance to polarity and is quite low for both models (as seen from the large gap of Gini Index decrease from polarity and the succeeding variables). The winner thematic model B1 produce of topic 17, length of the review and topic 19 as the outcomes of succeeding important predictors after polarity (ATU) score. Topic 17 and 19 can be viewed as the unsupervised LDA version of **PEOU** and LDA generated **ATU** respectively based on the terms that make up these two topics (table 5b). The next important predictor resulted from this model is topic 16, which revolves around the topic of **monetization** of the apps and its paid features. Topic 2 appears to contain the topic of **app version**, even though it is not very straightforward. Topic 11, 7, 8 can be seen as the unsupervised LDA version of **perceived_usefulness**, **vividness**, and **perceived_interactivity** respectively. It can be concluded, although this model wins over the concept-based approach of TAM via semi-supervised LDA, the most important features in determining the rating of the app consists of mostly TAM variable with a different terms constructs. On the other hand, **corona time** and its interction with PEOU and BI resulted from model C (*appendix A3 table 11*) do not result in any significant value (<0.05). This is contradicting oru expectation about the usability of VTO that improves during corona time.

7. Conclusion and Discussion

7.1 Discussion of Findings and Contribution

The systematic analysis that we have performed in the preceding chapters have led to the main findings of this research:

***RQ1:** How can text analysis method be applied in AR-driven YouCam Makeup apps' reviews to predict online rating?*

Firstly, by reviewing prior literatures in the field of immersive technology we obtained the first general impression of measurable features of immersive technology that is built upon our Technology Acceptance Model (TAM). Secondly, we followed the general procedures of text mining and text analytics techniques commonly used in marketing scholars to preprocess and extract features from the text data. In terms of the time consumed and practicability of the text preprocessing steps, string distance to correct misspell impose the most time-consuming process throughout the data preprocessing. Moreover, high precision in automatic misspell correction is almost unattainable due to the lack of computer's understanding in human's context. String distance is yet remain as a useful tool and a necessary practice for our dataset that is characterized by typography errors. It helped to recover many important words that may contain significant context about features of the YouCam Apps. This encourages future research to quantify the added value of various misspell correction techniques for a better clarity on its usefulness in information retrieval.

Moving forward, feature extraction is one of the most challenging tasks in this research as it required a careful consideration among various techniques that need to align with the goal of our research, resulted in the selection of sentiment analysis and LDA topic models. LDA topic model is mainly selected due to its ability to soft cluster that allows overlapping topics within a review window. Furthermore, it offers a great flexibility in integrating domain knowledge and statistical approach through its semi supervised version. Consequently, we are able to capture our conceptual understanding of TAM variables and adjust the degree of this expertise by various hyperparameter settings within the model in attempt to improve our topics' interpretation. Polarity score on the other hand is chosen due to its capacity in capturing valence shifters (amplifier, de-amplifier and negation) that can change the semantic meaning of the focal words and consequently, outperforms the traditional bag-of-words sentiment approach. This is considered as a more accurate representation of users' sentiment and it is reflected by its high node impurity in the variable importance plots, indicating the feature with the highest predictive power to the rating score.

By integrating all the corresponding techniques that results in the quantifiable input to our RF classification model, we found that our seeded LDA (model A) performed slightly poorer than unsupervised LDA (model B) in terms of accuracy. However, there are less overlapping of terms in model A resulting in better topic coherences than model B. Our winning model of B2 that is derived from pure statistical approach of unsupervised LDA performs best in terms of classification accuracy and Cohen's Kappa. It needs to be noted however, this technique requires more prerequisites steps of determining the number of topics and several hyperparameters a-priori that often are not a straightforward task without researchers' knowledge. Moreover, LDA has the downside of low interpretability; topics produced by this model contain many overlapping terms, resulting in vague topic definitions. Semi supervised LDA model, in comparison requires simpler prerequisites steps and higher interpretability due to the incorporation of the researcher's domain knowledge. Some might argue that a tradeoff of ±4% of accuracy for interpretability between seeded LDA and unsupervised LDA is worth achieving.

Nonetheless, we hope that this step-to-step analysis and the motivation behind the selection of every method have a considerable contribution to the marketing academia field. We hope to encourage future marketing research in exploiting online textual data that requires relatively less resources to collect than conducting an experiment and yet currently underutilized. Furthermore, it has the advantage of capturing the real time sentiment of users in the eWOM environment, allowing a deeper understanding and interpreting causes and effects of marketplace behavior. As human context remains as an important factor in text than other type of data sources, we especially recommend the predominant approach of seededLDA in the text mining application that is highly useful to integrate both human and machine learning's knowledge, resulting in a vastly superior insight as we have witnessed in the topics generated by model A.

In regard to managerial and industry implications, we hope to have shed a light in narrowing down the area of focus they need to put more attention into through the answer of our second research question:

***RQ2:** What are the most important drivers of the AR-driven YouCam Makeup apps in predicting online rating?*

Using topic models and sentiment analysis as our input to our supervised model (RF), we have found that polarity has the highest predictive power to rating score. This is expected, since it captures users' sentiment in using the product. Thus, it can be concluded that users' opinion and feelings about the adaption of VTO technology is strongly reflected in their style of writing that has a direct effect to the rating scale. Moreover, our winning model B2 also emphasizes the importance of PEOU (topic 17) as a second strongest predictor, that is composed over terms such as time, download, load, fix, uninstall, slow, and freeze as aspects that users perceived exhausting to their cognitive resources while using the AR technology. This LDA generated PEOU slightly differs from our own interpretation of this variable, that is composed over terms such as easy, crash, lag, quick, complicated, and stuck, that is derived from the subjectivity of seededLDA that depends partly in our own understanding of the TAM topics. This case is also applicable for topic 19, 11, 7, 8 as resulting topics of LDA generated topics; ATU, PU, vividness and PI in respect order. These variables are also placed in the top 5 strongest predictors in Model A, signaling managers be mindful for these aspects. Model A1 demonstrates the presence of enjoyment as a more important predictor than both PU and PEOU, confirming the expectation that YouCam apps delivers a higher hedonic value than utilitarian value. Although enjoyment is not reflected in our winning model w2, the mean decrease of Gini Index is higher (200) in the variable importance plot which would have held the third position in model B2. Another worth mentioning important features of YouCam Apps in addition to TAM cover the topics of monetization, paid features and app version (topic 2) concerning the aspects of value of money and its premium features. Corona time including its interaction with BI and PU on the other hand, does not result in any significant value, in oppose to our initial expectation regarding VTO's increase in usability during corona time.

7.2 Managerial Implications

By obtaining these key findings, we hope to inspire managers in the beauty industry that want to incorporate a successful adaptation of AR VTO by turning them into actionable insights within three focus areas; ease of use, visuals quality and the enjoyable aspects of VTO based on the order of our variable importance derived from the winning model B2 (and supporting model A1). These are the features that manager able to manipulate, unlike polarity and ATU (we cannot control users' sentiment and attitudes towards using the apps). Optimizing ease of use, visuals quality and enjoyable aspects of VTO on the other hand can have direct effects to polarity and ATU, as they meditate the presence of new technology to actual usage, demonstrated in the latent TAM model by Davis (1989).

Firstly, an effective implementation of VTO should not deprive users' cognitive resources. VTO, as a tool to facilitate product testing should be made simple and easy to navigate. Thus, managers need to work with the VTO developers in constant improvement of aspects such as time taken to load the apps or other VTO platforms and minimizing software bugs. These can be achieved by hiring skilled developers' and UX designers' team, regular test-driven development of the apps, and constant adaptations of the apps to the new versions of browsers and operating systems. Moreover, managers need to keep VTO simple by careful consideration on selected products to be integrated in the VTO. For instance, lipstick may be represented better in AR Try-On than a foundation, due to the more vibrant colors in which users could really notice the differences with one another. These efforts will consequently improve the visual quality of VTO, our second area of focus that is comprised of aspects of immersion and vividness. It is important that developers achieve the realness of the generated pictures that mimic the experience of real product testing.

Lastly, we recommend managers and developers to provide enjoyable and convenient shopping experience of VTO by allowing interactivity and sufficient options of product testing. This hedonic aspect of VTO characterizes modern shoppers that value technology integration of a brand and tend to promote them over social media and communities. As a result, customers that were not familiar with the brand might visit the website just to experiment with the AR technology. This increase in connectivity has a positive influence on brand awareness and loyalty by generating a sense of enjoyment and a greater novelty compared to web-based products presentation. This benefits managers by the expansion of word of mouth marketing, allowing them to position and target the products more efficiently by replacing the needs of unnecessary marketing spending. By following these recommendations, we hope to benefit the beauty industry in encouraging purchase intention through product testing and attracting new customers that are in the beginning of their customer journey.

7.3 Limitations & Future Research

Our research imposes some limitations. Firstly, semi supervised model such as seededLDA comes with the way researcher subjectively defines the topics. In most cases the choice of seed words per topic appears to be somewhat arbitrary. Intra-judge reliability, in which multiple researchers combine their expertise in defining topics can minimize the subjectivity and strengthen the topic coherency. Thus, topics generation process requires a lot of consideration that align with the researcher's goals and the abundancy of research resources. Moreover, the area of semi supervised LDA provides a potential exploration due to its usefulness and yet, not so many resources found in its useability; the SeededLDA package in R for instance, is a relatively new addition for the users' environment and there are not many resources regarding tweaking the combination of optimal hyperparameters.

Moreover, our findings on important features do not capture the directionality of our topics, that is derived from black box models such as LDA and RF. Black box algorithms are often complex and do not provide such a straightforward interpretation such linear regression. Additional black box interpreter techniques such as partial dependency plot, Shapley value or Lime method could help in uncovering the directionalities of our features. Lastly, our features, with the exception of polarity score, use a bag-of-word approach that do not capture the context in which features were mentioned in the text. For instance, we could not relate the users' sentiment towards each of the feature; trigram level of words such as "is not useful" have different meaning than the single word of "useful". Methods such as Word2vec or word embedding may tackle this limitation by mapping each word to a vector of latent dimensions or the context in which each word appears (Berger, et al., 2019). Understanding the context of human language remains as the biggest challenge in the field of data science and offers a great room for research. Nonetheless, we hope that this research could be a starting point to encourage the use of text data, semi-supervised NLP techniques, and provide frameworks for the industry to focus their resources in aspects that matter for a successful AR driven technology.

References

1. Accenture. (2014). Life on the digital edge: How augmented reality can enhance customer experience and drive growth. Retrieved April 2, 2017, from https://www.accenture.com/t20150521T005730__w__/usen/_ acnmedia/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Dualpub_8/Accenture-Augmented-Reality-Customer-Experience-Drive-Growth.pdf
2. Balakrishnan, V., & Ethel, L. (2014). Stemming and Lemmatization: A comparison of retrieval performances. *Lecture Notes on Software Engineering*, *2*(3), 262-267. https://doi.org/10.7763/lnse.2014.v2.134
3. Beck, M., & Crié, D. (2016). virtually try it … I want it! Virtual Fitting Room: A tool to increase on-line and off-line exploratory behavior, patronage and purchase intentions. *Journal of Retailing and Consumer Services 40*. http://dx.doi.org/10.1016/j.jretconser.2016.08.006
4. Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., & Schweidel, D. A. (2019). Uniting the tribes: Using text for marketing insight. *Journal of Marketing*, *84*(1), 1-25. https://doi.org/10.1177/0022242919873106
5. Blagus, R., & Lusa, L., (2010). BMC Bioinf., 2010, 11, 523.
6. Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning research, 3(Jan):993–1022.
7. Bonnin, G. (2020). The roles of perceived risk, attractiveness of the online store and familiarity with AR in the influence of AR on patronage intention. *Journal of Retailing and Consumer Services*, *52*, 101938. https://doi.org/10.1016/j.jretconser.2019.101938
8. Brownlee, J. (2017). *Deep learning for natural language processing: Develop deep learning models for your natural language problems*. Machine Learning Mastery.
9. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Advances in neural information processing systems, pages 288–296.
10. Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939785
11. Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, *43*(3), 345-354. https://doi.org/10.1509/jmkr.43.3.345
12. Damerau, F., 1964. A Technique for Computer Detection and Correction of Spelling Errors. *Communications of the ACM*, **7**(3): 171-176.
13. Darling, W. M. (2011). A theoretical and practical implementation tutorial on topic modeling and gibbs sampling. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, pages 642–647.

14. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American society for information science, 41(6):391–407.
15. Deloitte. (2019). *Technology in the mid-market: Seizing opportunity*. https://www2.deloitte.com/content/dam/Deloitte/us/Documents/deloitte-private/us-private-technology-mid-market-report-2019.pdf
16. Dhaoui, C., Webster, C. M., & Tan, L. P. (2017). Social media sentiment analysis: Lexicon versus machine learning. *Journal of Consumer Marketing*, *34*(6), 480-488. https://doi.org/10.1108/jcm-03-2017-2141
17. Digiday. (2021). *More brands are looking to augmented reality product try ons to drive sales*. https://digiday.com/marketing/brands-are-looking-to-augmented-reality-product-try-ons/
18. DigitalBridge (2017), "*Augmented reality – changing the face of retail*", available at: http://digitalbridge.eu/download-our-new-report-augmented-reality-changing-the-face-of-retail/ (accessed 11 November 2017).
19. Follett, L., Geletta, S., & Laugerman, M. (2019). Quantifying risk associated with clinical trial termination: A text mining approach. *Information Processing & Management*, *56*(3), 516-525. https://doi.org/10.1016/j.ipm.2018.11.009
20. Goldman Sachs. (2016). *Virtual & Augmented Reality, Understanding the Race for the Next Computing Platform*. https://www.goldmansachs.com/insights/pages/technology-driving-innovation-folder/virtual-and-augmented-reality/report.pdf
21. Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. Proceedings of the National academy of Sciences, 101(suppl 1):5228–5235.
22. Heller, J., Chylinski, M., De Ruyter, K., Mahr, D., & Keeling, D. I. (2019). Let me imagine that for you: Transforming the retail frontline through augmenting customer mental imagery ability. *Journal of Retailing*, *95*(2), 94-114. https://doi.org/10.1016/j.jretai.2019.03.005
23. Hilken, T., Heller, J., Chylinski, M., Keeling, D. I., Mahr, D., & De Ruyter, K. (2018). Making omnichannel an augmented reality: The current and future state of the art. *Journal of Research in Interactive Marketing*, *12*(4), 509-523. https://doi.org/10.1108/jrim-01-2018-0023
24. Hopping, D. (2000). Technology in retail. *Technology in Society*, *22*(1), 63-74. https://doi.org/10.1016/s0160-791x(99)00042-1
25. Jagarlamudi, J., Daumé III, H., and Udupa, R. (2012). Incorporating lexical priors into topic models. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 204–213. Association for Computational Linguistics.
26. Javornik, A. (2014). Classifications of augmented reality uses in marketing. *2014 IEEE International Symposium on Mixed and Augmented Reality - Media, Art, Social Science, Humanities and Design (IMSAR-MASH'D)*. doi:10.1109/ismar-amh.2014.6935441

27. Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing Imbalanced data--recommendations for the use of performance metrics. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. https://doi.org/10.1109/acii.2013.47
28. Kashefi, O., Sharifi, M., & Minaie, B. (2012). A novel string distance metric for ranking Persian respelling suggestions. *Natural Language Engineering*, *19*(2), 259-284. https://doi.org/10.1017/s1351324912000186
29. Kestenbaum, R. (2019). *The Future of Retail In The Beauty Industry Will Be Very Different*. Forbes. https://www.forbes.com/sites/richardkestenbaum/2019/09/04/the-future-of-retail-in-the-beauty-industry-will-be-very-different/?sh=207e329a6c4f
30. Khan, A. T. (n.d.). *How technology is revolutionizing beauty ecommerce*. Entrepreneur. https://www.entrepreneur.com/article/358018
31. King, W. R., & He, J. (2006). A meta-analysis of the technology acceptance model. *Information & Management*, *43*(6), 740-755. https://doi.org/10.1016/j.im.2006.05.003
32. Kwartler, T. (2017). *Text mining in practice with R*. John Wiley & Sons.
33. Liarokapis, F. (2006). An exploration from virtual to augmented reality gaming. *Simulation & Gaming*, *37*(4), 507-533. https://doi.org/10.1177/1046878106293684
34. Loo, M. (2014). The stringdist package for approximate string matching. *The R Journal*, *6*(1), 111. https://doi.org/10.32614/rj-2014-011
35. L'Oréal. (2020, February 20). *L'Oréal group: Discovering ModiFace*. https://www.loreal.com/en/beauty-science-and-technology/beauty-tech/discovering-modiface/
36. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*.
37. Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics, pages 100–108. Association for Computational Linguistics.
38. Owyang, J. (2010). Disruptive Technology – The New Reality Will be Augmented. *Customer Relationship Management Magazine, 32*(2), 32-33.
39. Ozturkcan, S. (2020). Service innovation: Using augmented reality in the IKEA place app. *Journal of Information Technology Teaching Cases*, 204388692094711. https://doi.org/10.1177/2043886920947110
40. Pallavicini, F., Pepe, A., & Minissi, M. E. (2019). Gaming in virtual reality: What changes in terms of usability, emotional response and sense of presence compared to non-immersive video games? *Simulation & Gaming*, *50*(2), 136-159. https://doi.org/10.1177/1046878119831420
41. Papagiannis, H. (2020). How AR Is Redefining Retail in the Pandemic. *Harvard Business Revew*. https://hbr.org/2020/10/how-ar-is-redefining-retail-in-the-pandemic

42. Park, M., & Yoo, J. (2020). Effects of perceived interactivity of augmented reality on consumer responses: A mental imagery perspective. *Journal of Retailing and Consumer Services*, *52*, 101912. https://doi.org/10.1016/j.jretconser.2019.101912
43. Perfectcorp. (2020). *Beauty AR company and makeup AR technology platform*. https://www.perfectcorp.com/business
44. Pollock, J. and Zamora, A., 1983. Collection and Characterization of Spelling Errors in Scientific and Scholarly Text. *Journal of The American Society for Information Science*, **34**(1): 51-58.
45. Poushneh, A. (2018). Augmented reality in retail: A trade-off between user's control of access to personal information and augmentation quality. *Journal of Retailing and Consumer Services*, *41*, 169-176. https://doi.org/10.1016/j.jretconser.2017.12.010
46. Poushneh, A., & Vasquez-Parraga, A. Z. (2017). Discernible impact of augmented reality on retail customer's experience, satisfaction and willingness to buy. *Journal of Retailing and Consumer Services*, *34*, 229-234. https://doi.org/10.1016/j.jretconser.2016.10.005
47. Rauschnabel, P. A., Felix, R., & Hinsch, C. (2019). Augmented reality marketing: How mobile AR-apps can improve brands through inspiration. *Journal of Retailing and Consumer Services*, *49*, 43-53. https://doi.org/10.1016/j.jretconser.2019.03.004
48. Reisenbichler, M., & Reutterer, T. (2018). Topic modeling in marketing: Recent advances and research opportunities. *Journal of Business Economics*, *89*(3), 327-356. https://doi.org/10.1007/s11573-018-0915-7
49. Rese, A., Schreiber, S., & Baier, D. (2014). Technology acceptance modeling of augmented reality at the point of sale: Can surveys be replaced by an analysis of online reviews? *Journal of Retailing and Consumer Services*, *21*(5), 869-876. https://doi.org/10.1016/j.jretconser.2014.02.011
50. Robertson, T. S., & Gatignon, H. (1986). Competitive effects on technology diffusion. *Journal of Marketing*, *50*(3), 1. https://doi.org/10.2307/1251581
51. Steyvers, M. and Griffiths, T. (2007). Probabilistic topic models. Handbook of latent semantic analysis, 427(7):424–440.
52. Tan, M., Tsang, I. W., & Wang, L. (2013). Minimax sparse logistic regression for very high-dimensional feature selection. *IEEE Transactions on Neural Networks and Learning Systems*, *24*(10), 1609-1622. https://doi.org/10.1109/tnnls.2013.2263427
53. Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation. *Journal of Marketing Research*, *51*(4), 463-479. https://doi.org/10.1509/jmr.12.0106
54. Watanabe, K., & Zhou, Y. (2020). Theory-driven analysis of large corpora: Semisupervised topic classification of the UN speeches. *Social Science Computer Review*, 089443932090702. https://doi.org/10.1177/0894439320907027

55. Wedel, M., Bigné, E., & Zhang, J. (2020). Virtual and augmented reality: Advancing research in consumer marketing. *International Journal of Research in Marketing*, *37*(3), 443-465. https://doi.org/10.1016/j.ijresmar.2020.04.004
56. Yim, M. Y., Chu, S., & Sauer, P. L. (2017). Is augmented reality technology an effective tool for e-Commerce? An interactivity and vividness perspective. *Journal of Interactive Marketing*, *39*, 89-103. https://doi.org/10.1016/j.intmar.2017.04.001
57. Zhou, J., Lee, I., Thomas, B., Menassa, R., Farrant, A., and Sansome, A. (2015). In-Situ Support for Automotive Manufacturing Using Spatial Augmented Reality. International Journal of Virtual Reality, 11(1):33–41.

Appendices

A.1 Theoretical Groundwork

Table 7: *Summary of response in immersive technology used in past literatures (positive outcomes)*

Factor	Definition	Reference
Learning effectiveness	Improvements in learning processes and outcomes, including level of content knowledge, academic achievement, performance, skills, ability and others	Frank and Kapila, 2017∗, Ibáñez et al., 2016∗, Yoon et al., 2012∗, Loup-Escande et al., 2017∗, Cheng and Tsai, 2014∗
Learning engagement	Increase in the amount of time spent focusing on AR/VR, a higher frequency of interactions	Ke et al., 2016∗, Chang et al., 2014∗
Learning attitude	Improvement in attitudes towards learning materials after experiencing the AR/VR	Hsiao et al., 2012∗, Hwang et al., 2016∗
Task performance	Improvement in efficiency (i.e., less than average completion time for correct actions) and accuracy (i.e., than less average overall error rate/higher success rate for tasks)	Radkowski et al. (2015) Zhao et al., 2016∗, Munafo et al., 2017∗
Reduced disease symptoms	Reduction in disease symptoms (e.g., pain, psychological stress, and mental diseases)	Mountford et al., 2016∗, Mosso-Vázquez et al., 2014∗, Hoffman et al., 2014∗ Pallavicini et al., 2016∗, Loreto-Quijada et al., 2014∗
Intention to use	User's intention to use AR/VR	Huang et al., 2010∗, Wojciechowski and Cellary, 2013∗, Yilmaz, 2016∗, Lee, Chung et al., 2013∗

Figure 8: *Extended TAM Model (King and He, 2006)*

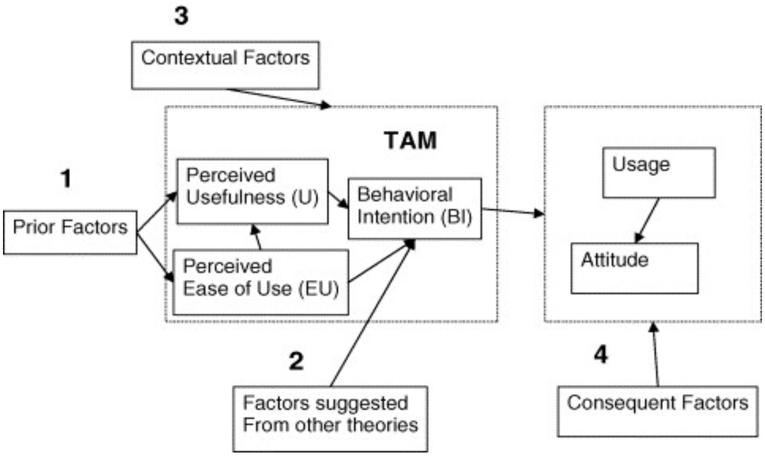


Figure 9: *Topic probability distribution of Model B (unsupervised LDA)*

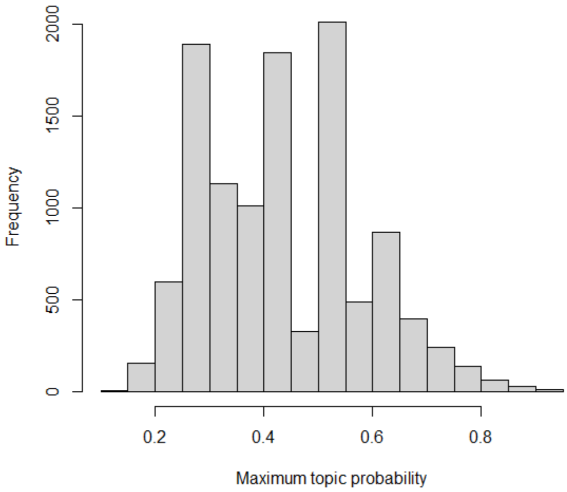
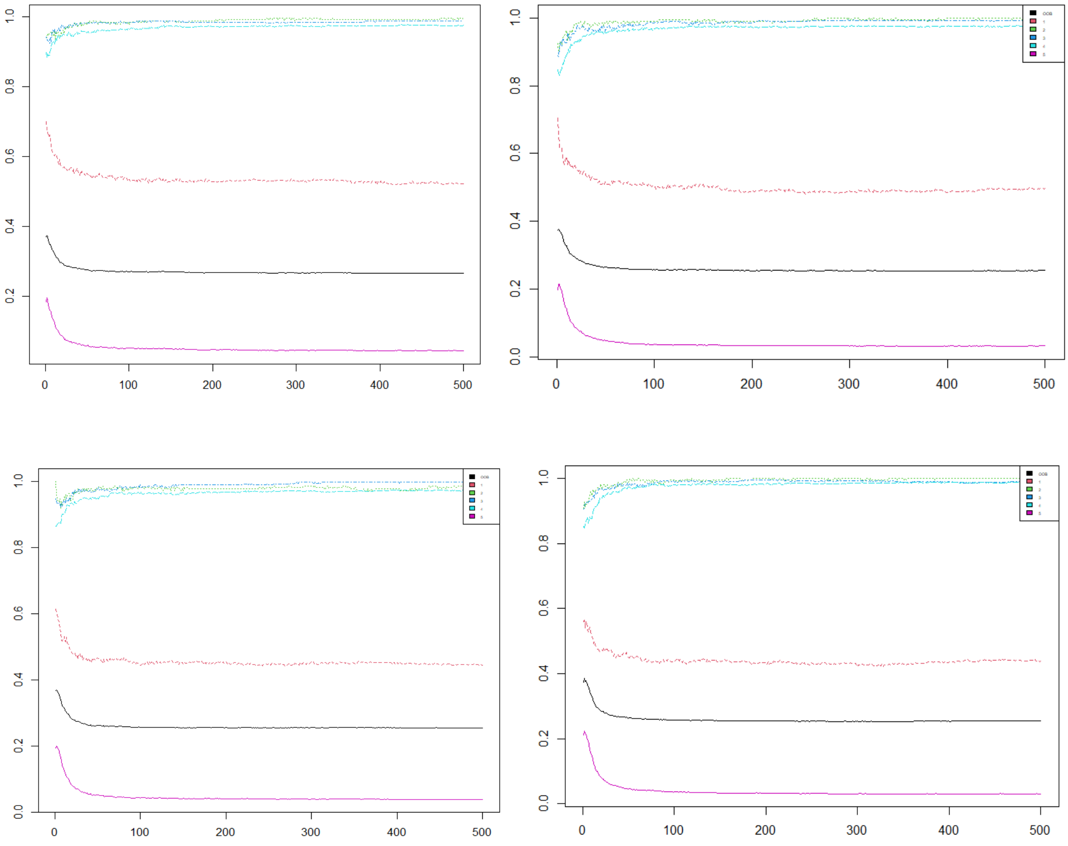


Figure 10: *Number of OOB Iterations per number of trees of RF Models (right to left: Model A1, A2, B1 and B2)*



A.2 Glimpse of the Datasets

Table 8: *The first 10 reviews*

No.	Review	Rating
1	i really like this app because the majority of the editing tools can be used without having to pay for the upgraded version. i also recommend this app because the results look natural not animated. even if a picture is completely changed through editing it still looks real. i really appreciate that because it is something hard to find.	5
2	this app is amazing i've been using it for years now and won't use anything else. but lately everytine i try to open the app it takes a very very long time to load. the screen just says updating data. it's so frustrating. i've had to uninstaller and reinstall twice now but it just keeps going right back to the same thing. other than that it's a great app. well it used to be i guess.	3
3	i've been using this app for years, and while i don't care that new features are behind a pay wall, i am furious that classic features have moved from free with ads to behind a pay wall. i have already downloaded a competitor. will not be using youcam again until they stop being greedy. i do not want your free trial, and i will not pay for a subscription for basic features available for free on every single other app.	1
4	love this makeup editing app i have been using it since . i do wish that they would keep all of the makeup features i miss some of the old brands that they used to use but that's not enough to make the app stars. best realistic makeup editing app. highly recommend for my gals out there if you don't have this app downloaded you download it now.	5
5	love this app, working really good, wanted to see how i would look with a certain hair colour and it looked realistic. i have colours in my hair so i tried the black dye to see if i would look like how i normally do and i did. would definitely recommend.	5
6	i love it. i've been using youcam for years and it just keeps improving. i'm kinda disappointed on some locked features that have to be suscribed to, but it remains the best makeup app.	5
7	this app is the #. the best app out there that lets you adjust the tones in the makeup section. it has a variety of colors and assesories to try out. thanks guys for this app.	5
8	amazing filters. this app is awsome. it lets you use cool make up for your face. i love how you can edit any photos you have in your camera roll and you the filters are wonderful.	5
9	i'm all the way around obsessed lets me try out looks before wasting makeup canastra better app realistic beautiful original fun worth every penny. would be stellar if they had a tutorial for the makeup looks they provide on here	5
10	i would give y'all a five star instead of a one but y'all have changed this a really bad were to we have to pay for it or we can't use it to fix our pictures change it back to we're it used to be free then y'all will get a five star, i use to love using this app now, i hate it cause of all the changes made to it.	1

***Table 9A:** Model A1 – data frame for the first 10 observations*

Rating	PI	Immersion	Vividness	Enjoyment	PEOU	PU	BI	Other	Corona Time	Trust	Polarity
5	0.42	0.19	0.04	0.04	0.12	0.04	0.12	0.04	post-Corona	4	0.96
3	0.45	0.05	0.05	0.05	0.05	0.05	0.25	0.05	post-Corona	0	0.58
1	0.32	0.03	0.09	0.32	0.03	0.09	0.09	0.03	post-Corona	1	0.88
5	0.14	0.14	0.14	0.05	0.05	0.05	0.41	0.05	post-Corona	1	0.57
5	0.06	0.17	0.39	0.06	0.06	0.06	0.17	0.06	post-Corona	2	2.25
5	0.32	0.05	0.05	0.05	0.05	0.23	0.14	0.14	post-Corona	1	0.58
5	0.28	0.06	0.28	0.06	0.06	0.06	0.06	0.17	post-Corona	0	0.58
5	0.05	0.05	0.05	0.05	0.15	0.55	0.05	0.05	post-Corona	1	1.44
5	0.04	0.11	0.11	0.18	0.04	0.25	0.04	0.25	post-Corona	1	1.25
1	0.25	0.05	0.05	0.05	0.45	0.05	0.05	0.05	post-Corona	2	0.28

***Table 9B:** Model A1 – data frame for the first 10 observations*

No.	Rating	PI	Immersion	Vividness	Enjoyment	PEOU	PU	BI	Other	Corona Time	Trust
1	5	0.42	0.19	0.04	0.04	0.12	0.04	0.12	0.04	post-Corona	4
2	3	0.45	0.05	0.05	0.05	0.05	0.05	0.25	0.05	post-Corona	0
3	1	0.32	0.03	0.09	0.32	0.03	0.09	0.09	0.03	post-Corona	1
4	5	0.14	0.14	0.14	0.05	0.05	0.05	0.41	0.05	post-Corona	1
5	5	0.06	0.17	0.39	0.06	0.06	0.06	0.17	0.06	post-Corona	2
6	5	0.32	0.05	0.05	0.05	0.05	0.23	0.14	0.14	post-Corona	1
7	5	0.28	0.06	0.28	0.06	0.06	0.06	0.06	0.17	post-Corona	0
8	5	0.05	0.05	0.05	0.05	0.15	0.55	0.05	0.05	post-Corona	1
9	5	0.04	0.11	0.11	0.18	0.04	0.25	0.04	0.25	post-Corona	1
10	1	0.25	0.05	0.05	0.05	0.45	0.05	0.05	0.05	post-Corona	2

No.	No of Words	Polarity	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Negative	Positive
1	18	0.96	0	1	0	0	3	0	0	0	6
2	14	0.58	0	2	0	0	0	0	1	0	0
3	27	0.88	1	1	2	0	1	0	0	2	2
4	20	0.57	0	0	0	0	1	0	0	0	2
5	11	2.25	0	1	0	0	2	1	1	1	4
6	11	0.58	1	0	2	1	1	1	0	2	3
7	11	0.58	0	0	0	0	0	0	0	0	0
8	12	1.44	0	0	0	0	2	0	1	0	3
9	15	1.25	0	1	1	1	2	1	0	1	5
10	15	0.28	2	2	2	3	3	2	0	2	3

Table 9C: *Model B1 – data frame for the first 10 observations*

No.	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	Rating	Corona_time	Trust	Polarity
1	0.01	0.23	0.01	0.01	0.01	0.01	0.01	0.06	0.23	0.06	0.23	0.01	0.01	0.12	0.01	0.01	0.01	0.01	0.01	0.01	5	post-Corona	4	0.96
2	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.78	0.01	0.01	0.08	3	post-Corona	0	0.58
3	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.77	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	1	post-Corona	1	0.88
4	0.01	0.14	0.01	0.01	0.14	0.01	0.14	0.01	0.07	0.01	0.01	0.01	0.01	0.34	0.07	0.01	0.01	0.01	0.01	0.01	5	post-Corona	1	0.57
5	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.76	0.01	0.01	0.01	0.01	0.01	0.09	0.01	0.01	0.01	0.01	0.01	0.01	5	post-Corona	2	2.25
6	0.01	0.01	0.01	0.01	0.01	0.01	0.28	0.01	0.28	0.01	0.19	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.10	5	post-Corona	1	0.58
7	0.01	0.12	0.01	0.12	0.34	0.01	0.01	0.12	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.12	0.01	0.01	0.01	5	post-Corona	0	0.58
8	0.01	0.01	0.01	0.08	0.01	0.01	0.16	0.01	0.01	0.01	0.16	0.08	0.01	0.01	0.01	0.01	0.08	0.01	0.32	0.01	5	post-Corona	1	1.44
9	0.22	0.08	0.01	0.15	0.01	0.15	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.22	0.01	0.01	0.08	0.01	0.01	0.01	5	post-Corona	1	1.25
10	0.01	0.32	0.01	0.19	0.01	0.01	0.01	0.01	0.38	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	1	post-Corona	2	0.28

Table 9D: *Model A1 – data frame for the first 10 observations*

No.	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17
1	0.01	0.23	0.01	0.01	0.01	0.01	0.01	0.06	0.23	0.06	0.23	0.01	0.01	0.12	0.01	0.01	0.01
2	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.78
3	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.77	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04
4	0.01	0.14	0.01	0.01	0.14	0.01	0.14	0.01	0.07	0.01	0.01	0.01	0.01	0.34	0.07	0.01	0.01
5	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.76	0.01	0.01	0.01	0.01	0.01	0.09	0.01	0.01	0.01
6	0.01	0.01	0.01	0.01	0.01	0.01	0.28	0.01	0.28	0.01	0.19	0.01	0.01	0.01	0.01	0.01	0.01
7	0.01	0.12	0.01	0.12	0.34	0.01	0.01	0.12	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.12
8	0.01	0.01	0.01	0.08	0.01	0.01	0.16	0.01	0.01	0.01	0.16	0.08	0.01	0.01	0.01	0.01	0.08
9	0.22	0.08	0.01	0.15	0.01	0.15	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.22	0.01	0.01	0.08
10	0.01	0.32	0.01	0.19	0.01	0.01	0.01	0.01	0.38	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

No.	V18	V19	V20	Rating	Corona Time	Trust	Polarity	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Negative	Positive	No of Words
1	0.01	0.01	0.01	5	post-Corona	4	0.96	0	1	0	0	3	0	0	0	6	18
2	0.01	0.01	0.08	3	post-Corona	0	0.58	0	2	0	0	0	0	1	0	0	14
3	0.00	0.00	0.00	1	post-Corona	1	0.88	1	1	2	0	1	0	0	2	2	27
4	0.01	0.01	0.01	5	post-Corona	1	0.57	0	0	0	0	1	0	0	0	2	20

5	0.01	0.01	0.01	5	post-Corona	2	2.25	0	1	0	0	2	1	1	1	4	11
6	0.01	0.01	0.10	5	post-Corona	1	0.58	1	0	2	1	1	1	0	2	3	11
7	0.01	0.01	0.01	5	post-Corona	0	0.58	0	0	0	0	0	0	0	0	0	11
8	0.01	0.32	0.01	5	post-Corona	1	1.44	0	0	0	0	2	0	1	0	3	12
9	0.01	0.01	0.01	5	post-Corona	1	1.25	0	1	1	1	2	1	0	1	5	15
10	0.01	0.01	0.01	1	post-Corona	2	0.28	2	2	2	3	3	2	0	2	3	15

A.3 Supporting Results

Table 10: *Confusion Metrics of all models*

<table>
<tr><td colspan="6"><b>Model A1</b>
<i>(seededLDA + Trust + Polarity +Corona + Time)</i></td></tr>
<tr><td colspan="6"><b>Overall Statistics</b>
Accuracy : 0.7363
95% CI : (0.7176, 0.7545)
No Information Rate : 0.7149
P-Value [Acc > NIR] : 0.2578
Kappa : 0.2578
Mcnemar’s Test P-Value : <2e-16</td></tr>
<tr><td>Sensitivity</td><td>0.45</td><td>0.04</td><td>0.02</td><td>0.01</td><td>0.96</td></tr>
<tr><td>Specificity</td><td>0.95</td><td>0.99</td><td>0.99</td><td>0.99</td><td>0.30</td></tr>
<tr><td>Pos Pred Value</td><td>0.52</td><td>0.13</td><td>0.13</td><td>0.14</td><td>0.78</td></tr>
<tr><td>Neg Pred Value</td><td>0.94</td><td>0.98</td><td>0.96</td><td>0.88</td><td>0.76</td></tr>
<tr><td>Prevalence</td><td>0.10</td><td>0.02</td><td>0.04</td><td>0.12</td><td>0.71</td></tr>
<tr><td>Detection Rate</td><td>0.04</td><td>0.00</td><td>0.00</td><td>0.00</td><td>0.69</td></tr>
<tr><td>Detection Prevalence</td><td>0.09</td><td>0.01</td><td>0.01</td><td>0.01</td><td>0.89</td></tr>
<tr><td>Balanced Accuracy</td><td>0.70</td><td>0.52</td><td>0.51</td><td>0.50</td><td>0.63</td></tr>
<tr><td colspan="6"><b>Model A2</b>
<i>(seededLDA + NRC Emotions + polarity + Corona Time + No. of words)</i></td></tr>
<tr><td colspan="6"><b>Overall Statistics</b>
Accuracy : 0.7377
95% CI : (0.7189 0.7558)
No Information Rate : 0.7028
P-Value [Acc > NIR] : 0.000143
Kappa : 0.2702
Mcnemar’s Test P-Value : <2.2e-16</td></tr>
<tr><td>Sensitivity</td><td>0.50</td><td>0.01</td><td>0.01</td><td>0.02</td><td>0.97</td></tr>
<tr><td>Specificity</td><td>0.96</td><td>1.00</td><td>1.00</td><td>0.99</td><td>0.28</td></tr>
<tr><td>Pos Pred Value</td><td>0.57</td><td>0.17</td><td>0.13</td><td>0.33</td><td>0.76</td></tr>
<tr><td>Neg Pred Value</td><td>0.95</td><td>0.97</td><td>0.96</td><td>0.88</td><td>0.82</td></tr>
<tr><td>Prevalence</td><td>0.10</td><td>0.03</td><td>0.04</td><td>0.13</td><td>0.70</td></tr>
<tr><td>Detection Rate</td><td>0.05</td><td>0.00</td><td>0.00</td><td>0.00</td><td>0.68</td></tr>
<tr><td>Detection Prevalence</td><td>0.09</td><td>0.00</td><td>0.00</td><td>0.01</td><td>0.90</td></tr>
<tr><td>Balanced Accuracy</td><td>0.73</td><td>0.51</td><td>0.50</td><td>0.51</td><td>0.63</td></tr>
<tr><td colspan="6"><b>Model B1</b>
<i>(LDA + Trust + Polarity + Corona Time)</i></td></tr>
<tr><td colspan="6"><b>Overall Statistics</b>
Accuracy : 0.7525</td></tr>
</table>

<table>
<tr><td colspan="6">95% CI : (0.7340, 0.7702)
No Information Rate : 0.7198
P-Value [Acc > NIR] : 0.0002752
Kappa : 0.3077
Mcnemar’s Test P-Value : <2.2e-16</td></tr>
<tr><td>Sensitivity</td><td>0.57</td><td>0.00</td><td>0.02</td><td>0.03</td><td>0.97</td></tr>
<tr><td>Specificity</td><td>0.95</td><td>1.00</td><td>0.99</td><td>0.98</td><td>0.35</td></tr>
<tr><td>Pos Pred Value</td><td>0.54</td><td>0.00</td><td>0.15</td><td>0.19</td><td>0.79</td></tr>
<tr><td>Neg Pred Value</td><td>0.96</td><td>0.97</td><td>0.96</td><td>0.88</td><td>0.80</td></tr>
<tr><td>Prevalence</td><td>0.09</td><td>0.03</td><td>0.04</td><td>0.12</td><td>0.72</td></tr>
<tr><td>Detection Rate</td><td>0.05</td><td>0.00</td><td>0.00</td><td>0.00</td><td>0.69</td></tr>
<tr><td>Detection Prevalence</td><td>0.10</td><td>0.00</td><td>0.01</td><td>0.02</td><td>0.88</td></tr>
<tr><td>Balanced Accuracy</td><td>0.76</td><td>0.50</td><td>0.51</td><td>0.51</td><td>0.66</td></tr>
<tr><td colspan="6"><b>Model B2</b>
<i>(LDA + NRC Emotions + polarity + Corona Time + No. of words)</i></td></tr>
<tr><td colspan="6"><b>Overall Statistics</b>

Accuracy : 0.7775
95% CI : (0.7597, 0.7946)
No Information Rate : 0.7381
P-Value [Acc > NIR] : 9.327e-06
Kappa : 0.3174
Mcnemar’s Test P-Value : NA</td></tr>
<tr><td>Sensitivity</td><td>0.58</td><td>0.00</td><td>0.00</td><td>0.02</td><td>0.98</td></tr>
<tr><td>Specificity</td><td>0.96</td><td>1.00</td><td>1.00</td><td>0.99</td><td>0.31</td></tr>
<tr><td>Pos Pred Value</td><td>0.60</td><td>NA</td><td>0.00</td><td>0.38</td><td>0.80</td></tr>
<tr><td>Neg Pred Value</td><td>0.96</td><td>0.98</td><td>0.97</td><td>0.88</td><td>0.82</td></tr>
<tr><td>Prevalence</td><td>0.09</td><td>0.02</td><td>0.03</td><td>0.12</td><td>0.74</td></tr>
<tr><td>Detection Rate</td><td>0.05</td><td>0.00</td><td>0.00</td><td>0.00</td><td>0.72</td></tr>
<tr><td>Detection Prevalence</td><td>0.09</td><td>0.00</td><td>0.00</td><td>0.01</td><td>0.90</td></tr>
<tr><td>Balanced Accuracy</td><td>0.77</td><td>0.50</td><td>0.50</td><td>0.51</td><td>0.64</td></tr>
<tr><td colspan="6"><b>Model C</b>
<i>(Model A + Corona*BI + Corona*PU + PU*PEOU)</i></td></tr>
<tr><td colspan="6">Accuracy : 0.7489
95% CI : (0.7304, 0.7668)
No Information Rate : 0.7189
P-Value [Acc > NIR] : 0.0007835
Kappa : 0.2465
Mcnemar’s Test P-Value : NA</td></tr>
<tr><td>Sensitivity</td><td>0.48</td><td>0.00</td><td>0.00</td><td>0.00</td><td>0.98</td></tr>
<tr><td>Specificity</td><td>0.96</td><td>1.00</td><td>1.00</td><td>1.00</td><td>0.25</td></tr>
<tr><td>Pos Pred Value</td><td>0.53</td><td>NA</td><td>NA</td><td>0.00</td><td>0.77</td></tr>
<tr><td>Neg Pred Value</td><td>0.95</td><td>0.97</td><td>0.96</td><td>0.88</td><td>0.83</td></tr>
<tr><td>Prevalence</td><td>0.09</td><td>0.03</td><td>0.04</td><td>0.12</td><td>0.72</td></tr>
</table>

Detection Rate	0.04	0.00	0.00	0.00	0.71
Detection Prevalence	0.08	0.00	0.00	0.00	0.92
Balanced Accuracy	0.72	0.50	0.50	0.50	0.61

Table 11: *Summary of Model C* (Multinomial Linear Regression)*

Residual Deviance: 14014.08

AIC: *14134.08*

<u>Coefficients:</u>

Class	*Intercept*	*PI*	*Immersion*	*Vividness*	*Enjoyment*	*PEOU*	*PU*
2	-1.50	0.74	0.58	2.04	-1.99	0.89	0.56
3	-1.20	0.32	2.75	3.73	-2.72	0.70	-2.63
4	-0.38	-2.63	2.95	3.21	-3.77	3.22	1.91
5	0.82	-3.60	3.72	1.56	-3.70	3.65	2.51
BI	***Other***	***Corona Time***	***Trust***	***Polarity***	***BI*Corona Time***	***PU*Corona Time***	***PEOU*PU***
-1.98	-2.34	0.09	0.23	0.11	0.68	0.06	8.17
-3.93	0.59	-0.37	0.31	0.65	1.80	3.55	3.38
-5.86	0.59	0.06	0.24	1.57	1.01	0.16	-1.23
-4.98	1.65	0.20	-0.01	2.13	0.31	0.38	-0.84

<u>Standard Errors:</u>

Class	*Intercept*	*PI*	*Immersion*	*Vividness*	*Enjoyment*	*PEOU*	*PU*
2	-0.33	0.78	1.20	0.97	0.80	1.73	2.90

3	*0.29*	*0.70*	*0.97*	*0.82*	*0.75*	*1.56*	*2.52*
4	*0.23*	*0.60*	*0.76*	*0.68*	*0.58*	*1.18*	*1.95*
5	*0.20*	*0.51*	*0.69*	*0.62*	*0.47*	*1.08*	*1.74*
BI	***Other***	***Corona Time***	***Trust***	***Polarity***	***BI*Corona Time***	***PU*Corona Time***	***PEOU*PU***
1.59	*1.18*	*0.35*	*0.09*	*0.12*	*1.15*	*2.04*	*18.92*
1.52	*0.88*	*0.30*	*0.08*	*0.10*	*1.09*	*1.66*	*16.55*
1.35	*0.68*	*0.25*	*0.06*	*0.08*	*0.99*	*1.37*	*13.24*
1.01	*0.59*	*0.21*	*0.061*	*0.07*	*0.75*	*1.23*	*12.28*

P-Value:

Class	***Intercept***	***PI***	***Immersion***	***Vividness***	***Enjoyment***	***PEOU***	***PU***
2	*7.19e-06*	*3.41e-01*	*6.26e-01*	*3.62e-02*	*1.27e-02*	*0.61*	*0.85*
3	*3.29e-05*	*6.51e-01*	*4.41e-03*	*5.52e-06*	*2.64e-02*	*0.65*	*0.30*
4	*9.84e-02*	*1.08e-05*	*1.14e-04*	*2.26-e06*	*9.05e-11*	*0.01*	*0.32*
5	*6.14e-05*	*1.36e-12*	*5.58e-08*	*1.18e-02*	*5.33e-15*	*0.00*	*0.15*
BI	***Other***	***Corona Time***	***Trust***	***Polarity***	***BI*Corona Time***	***PU*Corona Time***	***PEOU*PU***
2.14e-01	*0.05*	*0.81*	*9,78e-03*	*3.65e-01*	*0.55*	*0.98*	*0.67*
9,82e-03	*0.51*	*0.21*	*3.97e-05*	*7.27e-11*	*0.10*	*0.03*	*0.84*
1.34e-05	*0.38*	*0.80*	*1.51e-04*	*0.00*	*0.31*	*0.91*	*0.93*
8.48e-07	*0.00*	*0.34*	*8.57e-01*	*0.00*	*0.68*	*0.75*	*0.95*