Consumer Engagement on social media

Allan M

May 2, 2022

Abstract

THE ABSTRACT LATER.

Table of Contents

1 INTRODUCTION

1.1 Topic introduction and Motivation for Research

As of December 2021, out of about 7.8, about 4.9 billion people (62%) are on social media (ITU 2021) and on average, people spend 145 minutes daily, on social media (Statista1, 2021). The growth and ubiquity of social media, has led to the emergency and potency of social media influencers (hereafter called influencers). Influencers are people that have amassed a large number of followers on social media, either due to their career or due to producing engaging content e.g., Videos, Photographs, Blogs, etc. Partnership with influencers provides firms, the opportunity to adapt to this new trend and implement innovative marketing strategies to enhance brand value (Hughes et al. 2019).

Influencer Marketing is a growing industry, and is estimated to be a U$ 38-billion-dollar industry by 2024, and 75% of Marketers in the US, are reportedly using influencers for their brands (ANA 2020). Influencers are on every social networking site (SNS) e.g., YouTube, Facebook, Instagram, LinkedIn and more recently TikTok. However, Instagram is the most popular SNS for Influencer marketing, due to its photo and video sharing facilities (Haenlein et al. 2020). For example, on Instagram, one person alone, the football player Cristiano Ronaldo (@cristiano, has 401 million followers (Statista 2, 2021). That is more than the USA's total population of 331 million people (US census 2020).

Indeed, the giant sportswear manufacturer Nike, in 2016, signed a U$-billion-dollar-contract, with Cristiano Ronaldo, to market Nike's products. In 2016 alone, Cristiano Ronaldo made 1703 posts (21.7% of those posts, promoting Nike, through a video, photo, tweet, etc.). This led to an all combined 2.25 billion: comments, likes, shares, views, re-tweets, etc., on Facebook, Instagram, Twitter, etc. Hence, garnering U$ 474 million in value for Nike. (Forbes, 2016).

Furthermore, on social media and in concert with influencers, consumers interact with each other, and therefore influence each other, and propagate the brand message to others, through word of mouth (WOM). Research indicates that this has a higher engagement rate and a significantly higher response elasticity, compared to traditional marketing, where firms engage consumers directly (Trusov et al. 2009).

1.2 Problem Statement and Research Questions

Despite the rapid growth of influencer marketing, there is still a dearth of literature to delineate the industry. Hudders et al.(2020, p.7), conducted a systematic search of peer-reviewed academic papers on Influencer Marketing, from the Scopus data base. From their findings, only 154 papers were found, published between 2011 and April 2020. In the literature, there is a tendency to lump together influencers as one group. However, researchers (Campbell and Farrell 2020; Haenlein

et al.2020; Hudders et al.2020) indicate that, there are generally four categories of Influencers, namely: (1) Mega-Influencers ($\simeq$ 1 million + followers), (2) Macro-Influencers ($\simeq$ 100,000 - 1,000,000 followers), (3) Micro-Influencers ($\simeq$ 10,000 - 100,000 followers) and (4) Nano-Influencers ($\simeq$ 1,000 + followers). Furthermore, the research is either a brand-influencer, or influencer-consumer study (not the three parties together). Moreover, a number of the research studies are qualitative/surveys, that are typically based on relatively small and limited datasets, compared to online user generated content (UGC).UGC provide a rich source of data, to extract multiple dimensions of consumer engagement(Tirunillai and Tellis 2014).

For many SMEs (small-medium enterprises), with or without influencer marketing agencies, the problem is: (1) how to identify a competent influencer to partner with? (2) How to determine the measurement metric to measure engagement? (ANA 2020). To the best of my knowledge, no past research, has yet comparatively examined, consumer engagement on Instagram, by influencers in a brand-sponsored campaign. This therefore poses a research gap and hence, this paper attempts to contribute in filling that research gap. Specifically, this research tackles three research questions:

(1) Are there peculiar themes or topics, in consumers' responses (comments), to Influencers' engagement, and how can they be classified?

(2) Comparatively, are there differences in engagement rates, between Nano, Micro and Macro-Influencers on Instagram?

(3) How can firms strategically partner with Influencers, to benefit the brand's performance?

To answer these research questions, I use data crawled from Instagram. The dataset is obtained from brand-sponsored advertising campaigns. For each comment and like enlisted, I measure and operationalize engagement. I also apply topic modelling and word embeddings to ascertain the issues that are pertinent to the consumers, in response to influencers' posts. Finally, I build a predictive model that encompasses different features (topics,unigrams, bigrams, sentiments, influencer category) to predict consumer engagement.

1.3 Academic Relevance

This research makes a contribution to two streams of literature i.e., consumer engagement and influencer marketing on social media. Firstly, on consumer engagement, researchers: Hollebeek et al. 2019; Harmeling et al. 2017; Vivek et al. 2012, all conceptualize consumer engagement as a theoretical construct, however none of them, adduces a concrete measurement metric, to measure consumer engagement. Kumar et al.(2010), argue that consumers create value, and value can be measured. Hence, they examined consumer engagement value and suggested a measurement metric, consumer engagement value (CEV), that is also composed of consumer influence value (CIV). In line with Kumar et al.(2010), this paper advances that, by incorporating various variables in the

measurement of engagement, e.g., influencer category, hashtags, @, and emojis. This gives a better understanding of the drivers of consumer engagement on social media.

Secondly, Hudders et al.(2020) call for further research on, influencers on Instagram, particularly the different influencer categories i.e., Mega, Macro, Micro and Nano. Pursuant to that call, this research comparatively examines three influencer categories: Macro, Micro, and Nano influencers. The findings are that, Nano influencers have a higher engagement rate and are relatively more effective in engaging consumers on Instagram, compared to Micro and Macro influencers. This is line with the findings of Kay et el.(2020) and Park et al.(2021), who also find that, due to the perceived authenticity and more intimate relationships with lesser followers, Micro influencers are more effective and persuasive compared to Macro influencers.

Thirdly, Lou and Yuan. (2019); De Vierman et al. (2017); Evans et al. (2017), apply qualitative research/surveys, to examine influencers on Instagram. This research utilizes data that is, User Generated Content (UGC). Compared to surveys, UGC provide rich data sources, from which, multiple dimensions of Influencer Marketing and Consumer Engagement, are examined, i.e., influencer categories, consumer WOM (topics), consumers' emotions and sentiments (positive and negative consumer opinions). This enables a wider and deeper understanding of the peculiar factors that foster consumer engagement on social media (Tirunillai and Tellis 2014). Subsequently, prudent managerial policies can be ascertained and implemented. Furthermore, using UGC and a relatively large and diverse dataset, minimizes the error from selection bias.

1.4 Managerial Relevance

The findings from this study have managerial implications. Firstly, this research finds Nano influencers to be more persuasive and effective in engaging consumers, compared to Micro and Macro influencers. Thus, this study recommends Nano influencers to brands for marketing their products, particularly in niche markets. Hughes et al. (2019, p.81) posit that, firms engage influencers for two main purposes: (1) to create brand awareness and (2) to enhance product trial/purchase. Notably, influencers with larger numbers of followers, command a greater appeal for brands to reach larger audiences (De Vierman et al.2017, Jin and Phua 2014). However, due to more intimate relations with lesser followers, Nano (and Micro) influencers command relatively higher engagement rates and are perceived to be more authentic (Park et al.2021, Kay et al.2020). Moreover, according to the cultural meaning transfer model (McCraken 1989), consumers transfer such perceptions (influencer authenticity) to the product endorsed. Campbell and Farrell (2020, p.476) posit that, the different influencer groups (Nano, Micro, Macro, Mega) have different skillsets and are geared for different strategic marketing purposes. Thus, brands can utilize Mega and Macro influencers with larger audiences for product awareness, and employ Nano and Macro influencers with lesser followers, for product purchase.

Secondly, this research suggests to managers, to leverage Macro influencers (i.e., influencers with larger followings), by emphasizing the brand's authenticity be incorporated in influencers' posts, during the marketing campaign. Morhart et al. (2015, p.206) posit that, brand authenticity arises from: "(1) indexical authenticity i.e., objective facts, (2) iconic authenticity i.e., subjective mental associations and (3) existential authenticity i.e., current brand motives". Thus, influencers in their brand advertising, can utilize indexical, iconic and existential cues such as: communication that emphasizes the brand's history and virtues. This engenders perceived brand authenticity and thus consumer emotional brand attachment, positive word of mouth and purchase intention (Ilicic and Webster 2016).

Thirdly, in this study, visual images consistently generate consumer engagement, this research proposes to managers, to utilize and emphasize visual rhetoric, to be strategically incorporated in influencers' posts, during brand marketing campaigns. Visual rhetoric entails various elements in an image e.g., colour, shape, luminance, etc, and the details of the content the image displays, e.g., texture, colour variation, symmetric or asymmetric object portrayal, etc (Machado et al.2015). Visuals i.e., photographs influence consumer emotions of pleasure and arousal (Bakalash and Riemer 2013). Pleasure entails satisfaction and joy (Holmqvist and Lunardo 2015), while arousal is the stimuli to be engaged in something (Belanche et al.2017, Bakker et al.2014). The Stimulus-Organism-Response(S-O-R) model (Mehrabian and Russell 1974) posits that, stimuli influence consumers' emotions (pleasure and arousal) and subsequently consumer behaviour i.e., engagement or purchase intention (Vieira 2013).

1.5 Structure of the Thesis

In the next chapter (chapter two), literature is reviewed on: influencers on social media, consumer engagement, measurement of consumer engagement and the methodologies applied in the literature to analyze text data, using Natural Language Processing (NLP)/ machine learning techniques. In chapter three, the methodologies that this research applies, are elaborated and illustrated, to visually depict the research process and aims to be achieved. In chapter four, data employed and the data preparation process is explained. In chapter five, the findings and results of this paper are discussed. Finally, in chapter six, recommendations are given, conclusions are drawn, limitations of this research are elucidated, and suggestions for future research are then given.

2 LITERATURE REVIEW

The aim of this paper is to investigate consumers' responses to influencers' posts (consumer engagement) on Instagram, using Natural Language Processing (NLP) i.e., machine learning techniques. Thus, this chapter is divided into three parts, section 2.1 tackles influencers on Instagram, section 2.2. is a literature review of consumer engagement on social media. Finally, section 2.3 is about measurement of consumer engagement on social media. This chapter is supplemented with a table of the Literature on NLP techniques applied in this research.

2.1 Influencers on Instagram

Instagram is a social networking site (SNS) founded in 2010, geared for sharing photos and short videos among its users (Lee et al.2015). As of April 2022, there are about 1.13 billion worldwide users of Instagram and on average, people spend about 28 minutes per day on the social media (Socialpilot 2021, Shopify 2021). 60% of users are between the age of 18 to 34 (Pew Research Centre 2021). Instagram is primarily utilized for photo sharing, e.g., more than 100 million photographs are uploaded on the SNS every day worldwide (Socialpilot 2022). Instagram is the most popular social media for influencer marketers due to it's visual image and short video facilities (ANA 2020; Haenlein et al.2020; De Vierman et al.2017). Majority of the influencers on Instagram have lesser followers i.e. Micro and Nano influencers (54%) and the average engagement rate is 2.2% (Shopify 2021). Increasingly, brands are leveraging influencers on Instagram to market their products, and in 2021, brand advertising on Instagram was estimated to be U$5.8 billion (Shopify 2021). Influencers defer considerably based on: (1) number of followers, (2) engagement rates and (3) the value proposition they offer marketers (Campbell and Farrell 2020). Thus, next is a brief review of influencer categories on Instagram, i.e., Celebrity, Mega, Macro, Micro, and Nano influencers.

Celebrity influencers have a million + followers on Instagram and they attained fame due to their careers e.g., as athletes, musicians, actors/actresses, etc, prior to being famous on social media, e.g., the football player Cristiano Ronaldo (@cristiano with 432 million followers) or the TV personality, Kim Kardashian (@kimkardashian with 303 million followers). Given their public recognition, brands utilize them, to appear in adverts endorsing products (McCraken 1989, p.310). They particularly promote luxury and high-end products (Bearden and Etzel 1982), are iconized by their followers and thus possess celebrity status and cultural capital that brands utilize to promote awareness (Kelting and Rice 2013). However, given their high number of followers, they lack personal and intimate connection, with followers and therefore have the lowest engagement rates on social media (Campbell and Farrell 2020).

Mega influencers also have a million + followers on Instagram, unlike celebrities, Mega influencers attained their fame based on the content they have produced and posted on social media. Their content tends to be more detailed e.g., blog posts, tutorial videos on YouTube. They are regarded

as experts within certain domains, e.g., Kylie Jenner is a cosmetics expert (@kyliecosmetics, with 25 million followers on Instagram). Being perceived as experts bestows on them credibility and this engenders persuasiveness and influences information impact on consumers (Bearden and Etzel 1982). Unlike Celebrities whose appeal is aspirational, consumers follow Mega influencers to seek expert knowledge, especially when they are interested in, or are to purchase high-involvement products e.g., personal computers, cars, etc. Experts often review the positive and negative sides of products in their industry and this makes them highly regarded. Thus, brands utilize Mega influencers for advertising that requires conversant knowledge and more detailed information about the functionality of the product (Mudambi and Schuff 2010).

Macro Influencers have between 100,000 to 1,000,000 followers, they command relatively higher engagement rates compared to Mega and Celebrity influencers (Campbell and Farrell 2020). Given their sizable audience with common interests in certain domains among their followers e.g., fitness, travel, food, etc, it is more ideal for brands to target such consumers compared to blanket online targeting, because consumers are more likely to respond to adverts placed within content, that is in their area of interest (Taylor 2009, Coulter 1998). Additionally, adverting brands are often encumbered by consumer advert skipping or blockage, as digital adverts are perceived to be inconvenient and untrustworthy (Cho and Cheon 2004). Thus, this enables firms advertising their products to reach otherwise difficult to reach audiences.

Micro influencers have between 10,000 to 100,000 followers and their followers are oftentimes geographically localised (Campbell and Farrell 2020). Compared to Mega and Macro , Micro influencers are usually more involved with social media, so as to engage their audience (perhaps to grow their audience as well) and often utilize facilities such as Instagram stories (videos), such videos are significant drivers of engagement and that creativity obtains more consumer attention than traditional adverts by brands (Pereira et al.2014). Moreover, research indicates that consumers respond positively to adverts placed within content they enjoy (Van Reijmersdal et al.2010)

Nano influencers have the smallest follower base (below 10,000). Their followers are more often acquaintances e.g., friends or members of a local community (Campbell and Farrell 2020). This organic reach enables personal accessibility and perceived authenticity (Lipsman et al.2012). Thus, their posts are perceived to be more authentic and persuasive (Kowalczyk and Pounders 2016). Nano influencers are particularly an effective means for brands to reach niche audiences in emerging trends with in the market (Campbell and Farrell 2020, Shopify 2021). In sum, each influencer category, offers brands, different value considerations. The combination of deeper audience knowledge, creativity, expertise and cultural capital, means that managers need to ensure a suitable match between the brand and the influencer or utilise the different influencer categories within the same marketing campaign (Fleck et al.2012).

Influencers' posts initiate and often times are the main driver of engagement on social media. Alba and Hutchinson (1987) show that consumers' (influencers') expertise or competence, manifests

inform of : creativity, knowledge within a certain domain, and the ability to utilize that knowledge, to persuade and galvanize others to a certain cause. An influencer's abilities and creativity through their posts, affects consumers' attitude to act or change (McCracken 1989) and attitude change leads to a change in behavioural intention e.g., further engagement or purchase of a product (Uribe et al.2016). On low-involvement media, such as on Instagram, affective tools are instrumental in eliciting pleasure and arousal (Bakalash and Riemer 2013). Thus, hedonic content in form of visual images and short videos are key, as a peripheral cue, to generate engagement on social media (Berger and Schwartz 2011).

This is in line with the Elaboration Likelihood Model (ELM) by Petty and Cacioppo (1986). Moreover, Berger and Milkman (2012) find that, hedonic content drives virality on social media. Additionally, evidence indicates that, engagement on Instagram is positively related with the hedonic content of the posts (Li and Xie 2020). This paper therefore hypothesizes that, an Instagram influencer's expertise and hedonic post, interact and lead to a higher level of engagement. In this study, an influencer's expertise is derived from the relatively high engagement a post generates and the sentiments expressed on the post, by the consumers.

2.2 Consumer Engagement on Social Media

Customer Engagement (CE) as a concept in the marketing academic discipline, has many definitions, e.g. Van Doorn et al.(2010) view CE from a behavioral perspective, Brodie et al.(2011) approach CE from a psychological point of view, Kumar et al.(2010) posit a value-based perspective, while Lemon and Verhoef(2016) define customer engagement as "consumers reaching out". This paper specifically focuses on indirect consumer engagement (consumer-to-consumer) which entails word-of-mouth (WOM), emotionality and trust (Harmelling et al.2017, Pansari and Kumar 2017).

Trust is an integral factor that fosters interactions on social media. Moorman et al. (1993) define trust as being ready to count on another person for exchange, in whom one has faith in. Garbarino and Johnson (1999) implemented a 5-point measurement scale for consumer trust and the authors show that, trust can be measured and manifests itself inform of: repeated purchases, referrals and positive WOM. Consumer engagement studies show that, trust amongst consumers, is crucial for interactions to flourish on social media, e.g., Racherla et al.(2012) examined consumers' trust in online product reviews, and found that, a consumer's reviews (content) and perceived reviewer's background similarity to other consumers, significantly enhance trust in that person, while Hollebeek (2011) argues that, trust creates a sense of belonging to an online community and therefore fosters WOM online.

The extant literature on consumer engagement posits that, both positive and negative emotions, are drivers of engagement on social media (Pansari and Kumar 2017, Santini et al.2020). Richins (1997) empirically measure consumption-related emotions, using a 4-likert scale, they show that

positive emotions include: joy, excitement surprise, while negative emotions include anger and sadness. Emotions are known to engender intense focus on a certain object, which then leads to certain consumer behaviour. Vivek et al.(2012) opine that, the emotional intensity and extent to which, consumers are engaged in online interactions, inevitably leads to certain online behaviour. Behaviour may include purchases, but oftentimes word-of-mouth is influenced by emotions. Thus, this paper hypothesizes emotions to be potential drivers of engagement in this research.

2.3 Measurement of Consumer Engagement on Social Media

In the Marketing Literature, value is largely defined monetarily, e.g., Gupta and Lehmann (2005) consider value, as the financial value that the firm attains from consumers transactions. However, Kumar et al.(2010) refute that notion, arguing that, it is insufficient and therefore undervalues or overvalues consumer engagement, and the value the firm attains. The authors propose a measurement metric namely: customer engagement value (CEV), that has four components: (1) customer life-time value (CLV), (2) customer referral value (CRV), (3) customer knowledge value (CKV) and (4) customer influence value (CIV).

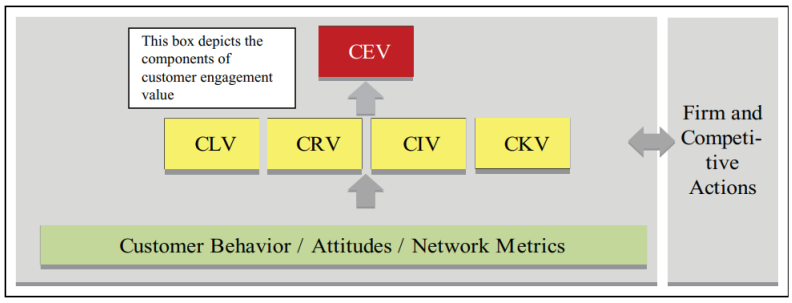


Figure 2.1: Conceptualizing and measuring CEV
Source: Kumar et al.(2010)

Kumar et al. (2010) argue that: Customer Life-time value, the present value of a consumer's future purchases, is futuristic and thus, it is not captured in value that traditionally, is based on current transactions. Customer referral value, stems from the firm's initiatives and incentives, such that, through the firm's engagement, some customers refer others to the firm's offering. Customer referrals reduce acquisition costs and increase the firm's future revenues (Ryu and Feick 2007).

Customer knowledge value, is attained through engaged consumers, who poses extensive product knowledge and consumer preferences. Through connected, knowledgeable and empowered consumer brand communities, firms gain feedback that is essential for product development (Prahalad and Ramaswamy 2004). Customer influence value is obtained through word of mouth (WOM) and consumer interactions on social media networks. In sum, Kumar et al. (2010) argue that, value from consumer engagement, has multiple dimensions, some of the value to the firm, accrues in the future and the suggested four components of CEV, interact with each other. This research builds

on the authors' measurement metric i.e customer influence value (CIV), to measure influencers' consumer engagement on social media.

Kumar et al.(2013) in their award-wining marketing research (MSI Practice Winner 2011-2012), measure engagement (comments and hashtags) on Facebook and Twitter, and the monetary impact of consumers' Word of Mouth (WOM) in generating sales. The authors create a metric, they call Customer Influence Effect (CIE), that measures the influence of a consumer on social media. They then link the influence to a monetary value (gain or loss), i.e., Customer Influence Value (CIV), that the firm (Hokey Pokey) obtains. From historical data on social media, they identify key individual influencers, incentivize them and through a hashtag campaign, influencers spread WOM on Facebook and Twitter. Finally, through special tracking software installed in the operations of the retailer (Hokey Pokey), and using advanced statistical methods, they gather user-level data and sales data. In the final analysis, the brand awareness of the firm increased by 49%, sales growth by 40% and ROI by 83%. Importantly, Kumar et al. (2013) emphasize WOM, i.e., consumer topics of discussion, during the campaign, from which they identified key consumers (influencers) and consumer preferences.

Hughes et al. (2019) examine consumer engagement by influencers on their Blogs and Facebook pages simultaneously, at different stages of the consumer purchase funnel (i.e., product awareness vs product trial). Their findings are that, blogger-expertise is significant on Blogs (high-involvement and less-distraction media), but not on Facebook (low-involvement and high-distraction media). Furthermore, hedonic-content posts, garner more engagement on Facebook, than on Blogs. In sum, blogger expertise, media type and post content (hedonic vs informational) interact differently to generate engagement. The authors measure engagement by the summation of comments and likes.

Operationalization of engagement (count of comments and likes) as researched by Hughes et al.(2019), Rooderkerk and Pauwels(2016), highlights the amount or level of engagement on social media. However, in that approach, the substance of engagement is not ascertained, and the distinction of the comment features as drivers of engagement is not established. This paper posits that, comment features can be identified, e.g. topics, sentiments(consumers' positive and negative opinions), emotions, etc. Essentially, the comments as a whole, drive engagement, through the sum of their individual parts. Thus, this research utilizes Natural Language Processing (NLP) to mine the comments and then operationalize engagement. This delineates the different features, and their significance in driving engagement on social media. Next is a table of Literature for NLP methods, this research utilized to build models to analyze consumer engagement.

2.3.1 NLP Literature Review - Data and Methods

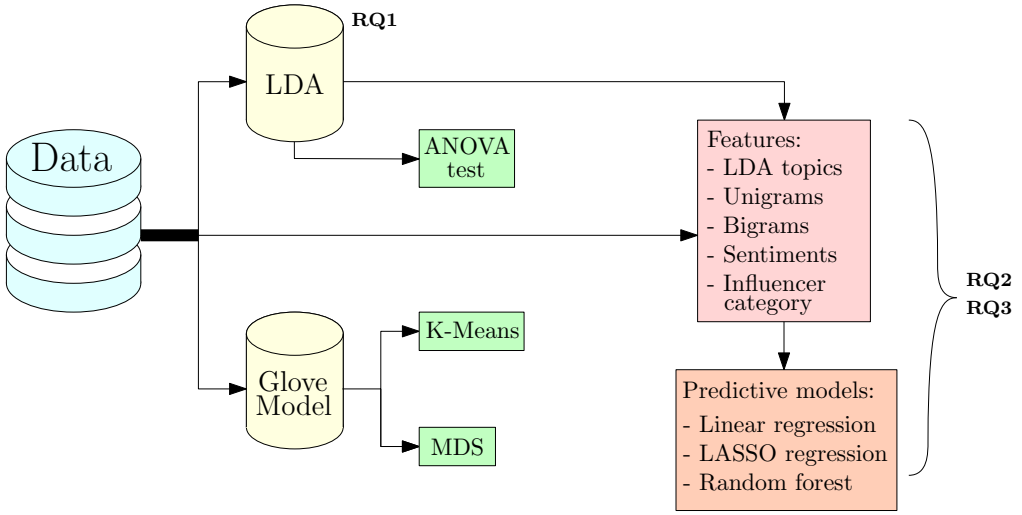
Research	Social Media	Data	Features / variables	Key independent variables	Unsupervised Learning Method / Comment Features	Supervised Learning Method	Dependent variables
Aldous et al. (2019)	Facebook, Instagram, Twitter, YouTube, Reddit	Comments, Likes	Topics, bigrams, sentiments, emoji, ?, !	Topics, bigrams	Latent Dirichlet Allocation	Random Forest, AdaBoost	Engagement
Lee et al. (2018)	Facebook	Comments, Likes	Question, blanck, emoticon, bigram, emotion	Question, blanck		Logistic Regression	Engagement
Jaakonmäki et al. (2017)	Instagram	Comments, Likes	Number of followers, posting time (a.m/p.m), emoji, words	Number of followers, posting time (a.m/p.m)		Lasso Regression	Engagement
Bakshi et al. (2014)	Instagram	Comments, Likes	Photographs, number of followers, words	Photographs, number of followers		Negative Binomial Regression	Engagement
Risch and Krestel (2020)	TheGuardian.com	Comments, Likes	Number of words, readability index, sentiment	Number of words, readability index		Logistic Regression, Convolutional Neural Network	Engagement
Suh et al. (2010)	Twitter	Tweets	URL, hashtag, mention (@username), follower	URL, hashtag	Principal Component Analysis	Generalized Linear Model	Engagement
Jamali and Rangwala (2009)	Digg	Comments	Number of words, number of comments	Number of words, number of comments		Decision Trees, Support Vector Machine (SVM)	Engagement
This research	Instagram	Comments, Likes	Influencer category, unigrams, topics, bigrams	Influencer category, bigrams	Latent Dirichlet Allocation, Glove Model	Linear Regression, Random Forest Regression, Lasso Regression	Engagement

3 METHODOLOGY

This chapter explains the methodologies applied in this research. The chapter is divided in three sections. Section 3.1 summarizes the methods and research process applied, to answer the 3 Research Questions(RQ) this paper tackles. This is supplemented with a visualization to illustrate the research process. Section 3.2 focuses on the unsupervised learning methods applied in this paper. Finally, section 3.3 is about supervised learning methods (predictive models) used in this research. This is also complemented with a visualization to depict the predictive models applied.

3.1 Research Process

Figure 3.1: Research Framework



Using Latent Dirichlet Allocation (LDA), coupled with ANOVA tests (to test for the optimal k number of topics), the underlying topics within the comments are detected, categorized and labelled. This answers research question (1): "Are there themes or topics, in consumers' responses (comments) to Influencers' posts and how can they be classified?"

Subsequently, topics and other comment features i.e., unigrams, bigrams, sentiments, influencer category, are extracted and applied in predictive models. Coupled with t-tests, this answers research question (2): "Comparatively, are there differences in engagement rates, between Nano, Micro and Macro Influencers on Instagram?"

Finally, for research question (3): "How can firms strategically partner with influencers, to benefit the brand's performance?" This is answered by the combination of the methods applied in this research.

3.2 Unsupervised Learning Methods

3.2.1 Latent Dirichlet Allocation (LDA)

Topic models are unsupervised machine learning techniques that detect topics from a collection of documents in a corpus. Topic models employ a global context approach, that ignores the order of words and documents in the corpus. Therefore, the ordering is exchangeable for documents and words. However, the underlying latent themes (topics) have a conditional distribution (Blei et al.2003).

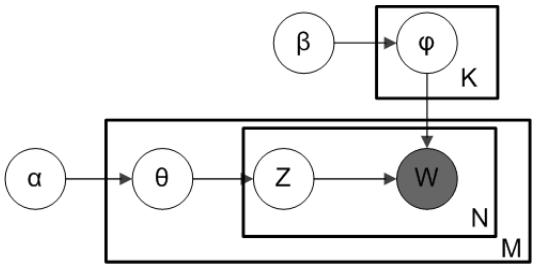
The underlying latent themes are based on the statistical properties of documents in the corpus. The topic model algorithms uncover the latent themes, summarize them into, categories of overarching topics, that can be meaningful to interpret and be labelled by the researcher. Such topics are potentially significant factors for consumer engagement. Therefore, in consumer engagement analysis, topics can be examined individually to ascertain their substance and significancy in a marketing campaign (Kumar et al.2013).

Topic models such as Latent Semantic Analysis or Indexing (LSI) analyse documents and uncover concepts and related terms in a corpus, however LSI has drawbacks, such as lacking robust probabilistic modelling. Probabilistic Latent Sematic Indexing (pLSI) is an improvement of LSI, in that, it models words as samples from a mixture model, whose components can be depicted as topics (Blei et al.2003). Still pLSI also has limitations, whereby, there is no generative probability modelling for the mixture proportions of topics, and with an increase in the size of the corpus, this leads to overfitting (Blei et al.2003).

Latent Dirichlet Allocation (Blei et al.2003) improves both LSI and pLSI, by combining the limitations of both LSI (no robust probabilistic modelling) and pLSI (no generative probabilistic modelling). LDA is a generative probabilistic topic model that detects topics from unstructured document collections in a corpus. LDA also holds a bag-of-words assumption, where document-and-word order is neglected. LDA is generative in a way that, documents are random mixtures over latent topics and each topic has a distribution over a fixed vocabulary of words. There are 3 steps in LDA's generative process:

1. Choose distribution over topics $\theta_i \sim \text{Dir}(\alpha)$ where $i \in (1, \ldots, M)$
2. Choose distribution over words $\varphi_k \sim \text{Dir}(\beta)$ where $k \in (1, \ldots, K)$
3. From each of the word positions i, j
   (a) Choose a topic $z_{i,j} \sim \text{Multinomial}(\theta_i)$
   (b) Choose a word $w_{i,j} \sim \text{Multinomial}(\varphi_{z,j})$

Figure 3.2: LDA Plate Notation



From the plate notation in Fig.3.2.

The plates (N for words, M for documents, K for topics) represent regeneration in the corpus (words-documents-topics) and the edges of plates indicate the conditional dependence between the variables. Hence, the three-level hierarchical structure. α and β are global parameters, α is the Dirichlet prior per-document topic distribution, β is the Dirichlet prior per-topic word distribution, θ is the topic distribution for a document, φ is the word distribution for a topic, z is a topic for a given word, and w is a particular word. N is the number of words in a document, M is the number of documents. W is highlighted and is the only observable variable, to detect the latent themes and therefore apply LDA, a researcher needs to determine the posterior distribution:

$$p(\theta, \varphi, z \mid w, \alpha, \beta) = \frac{(\theta, \varphi, z, w \mid \alpha, \beta)}{p(w \mid \alpha, \beta)}$$

The posterior distribution is intractable, due to the complexity of computing latent (hidden) variables. Because distributions have complicated non-linearities. Thus, statistical inference methods are used to estimate the posterior. Gibb's sampling and Variational Expectation Maximization (VEM), are alternative inference algorithms in R's package for topic models. This paper used VEM as it estimates α, however, both methods yield sufficient results. After obtaining the required parameters, I applied an ANOVA test to ascertain the optimal number of topics. Finally, the topics were obtained and labelled (obtained topics will be discussed further in chapter 5).

3.2.2 Global Vectors for Word Embeddings (Glove Model)

Glove Model (Pennington et al.2014) is a model for word embeddings, in which vector representations describe the meaning of words in a body of text or corpus. Unlike topic modelling that only takes a global approach, the Glove Model takes both a local context, where the meaning of a focal word is derived from the neighbouring words (e.g., a window size of z words before and after a focal word) and a global context, where training of the model is undertaken on a global word-word co-occurrence matrix.

The meaning of words as deduced from the embeddings(vectors), is obtained from the syntactic and semantic properties of the words. Thus, words from a similar category e.g., nouns, verbs,

prepositions, etc, have similar meaning(embedding). The syntactic aspect or local context of the words, is also captured in the embeddings, e.g., the sentence structure.

For example, the semantic properties of the word "man" are: he, male, king, etc, while "woman": she, female, queen, etc. Therefore, royalness can be inferred from the word arithmetic: king - man = queen - woman. Equally, with syntactic, e.g.: "today he is walking on the road". From word embeddings, "walking" is related to "road" (sentence structure: verb-object), but also "today" is related to "walking" (present tense).

Essentially, word embeddings reveal the similarity, relationship and context of the words. This enables an analysis in this research, that captures the subtleties and nuances of consumer comments. In contrast, topics are broad themes detected from the whole corpus. The juxtaposition of LDA and Glove Model can potentially reveal interesting insights.

By comparing topic modelling and word embeddings, this paper aims to ascertain which model performs better and is more interpretable to be applied in the prediction model. Secondly, given that the two models (LDA and Glovel Model) apply contrasting approaches (global vs local context approach), it is informative to know whether the topics detected by LDA and Glove Model word embeddings, are congruent or incongruent in terms of their results.

Furthermore, in LDA, term frequency-inverse document frequency(tf-idf), controls for word frequency, however the associated meaning between words is not captured. The Glove Model applies statistics that capture the relatedness among words, while also controlling for word co-occurrence frequency. Hence, obtaining more fine-grained word similarity and meaning in the body of text. This complements the analysis of this research, in that, the intrinsic and substantial meaning of consumer comments is captured.

In the final analysis, a deeper evaluation of consumer engagement is attained. Thus, word embeddings reveal the meaning of comment features that are related with other entities in consumer comments. So, consumers may respond to an influencer's post or other consumers' comments, not in isolation, but rather due to a connection with other linkages within and in the context of the interaction.

Notation for Glove Model

X is the global word-word co-occurrence matrix, that is decomposed into rows and columns of vectors, such that X_{ij} enumerates the co-occurrence frequency of words ij. $P_{ij} = P(j \mid i) = X_{ij}/X_i$ is the probability that word i occurs in the context of word j. X_{ik} tabulates the frequency word i appears with any word k.

For example, given 3 words (ijk), word i = pen, word j = paper, k = ink, or k = white, or k = stationery, or k = noise. For words related to pen but not related to paper, e.g., k = ink, the ratio

P_{ik}/P_{jk} will be high. Equally, for words related to paper but not pen e.g., k = white, the ratio P_{ik}/P_{jk} will be low. Finally, for words related to both pen and paper e.g., stationery, or words not related to both pen and paper e.g., noise, the ratio P_{ik}/P_{jk} will be near 1.

Compared to word probabilities, the odds-ratios are more informative, because they reveal and sort out words that are relevant(ink and white), from words that are irrelevant(stationery and noise). From the odds-ratios, the word embeddings/numerical vectors, are then obtained in the matrix. The vectors are points in a multidimensional dimensional space, and are the meaning or similarity between words. Thus, the cosine similarity of words is utilized to infer the underlying relationship between words.

In the global co-occurrence matrix X , the focal word and the context word, are arbitrary and therefore exchangeable, so $w \leftrightarrow \tilde{w}$ and also $X \leftrightarrow X^T$. Some words are more related, therefore have high co-occurrence frequencies (higher vector values), while others are not (low frequency and low vector entries), a bias term b_i is thus incorporated as a weight control for word frequency, $\tilde{b}_k$ is for symmetry, for the context word and $w_i^T \tilde{w}_k$ is a measure of fit between the words. X is logarithmic, so as to scale the distribution. X is a sparse matrix and a logarithm entry of 0, makes it undefined and unstable. Thus, an additive shift is added to maintain its sparsity and stability.

$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log(1 + X_{ik})$$

However, the drawback of the above equation and other Latent Semantic Analysis (LSA) models is that, all co-occurrences are weighted equally, including those that infrequently co-occur or never co-occur (Pennington et al.2014). Thus, given a vocabulary size (V) in the corpus, Pennington et al.(2014) suggest a least squares weighting function $f(X_{ij})$, where frequent and infrequent co-occurrences are not over weighted.

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2$$

where

$$f(x) = \begin{cases} (x/x_{\max})^\alpha & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases}$$

Pennington et al.(2014, p.4) propose a weighting function f with $\alpha = 3/4$ and a cutoff point of $x_{\max} = 100$, at which point, there no additional weights for frequency and the function flattens. From the authors' empirical finding, that significantly improves the model. This research therefore followed the same approach. After obtaining the word embeddings global co-occurence matrix, and following Ahmad and Amin (2016), who also used K-Means on the Glove Model, K-Means was applied with k = the obtained k for LDA topics. Essentially, this was done, to compare both models (LDA and Glove Model), based on the same number of topics and clusters. The results for both models will be discussed in chapter 5.

3.3 Supervised Learning Methods

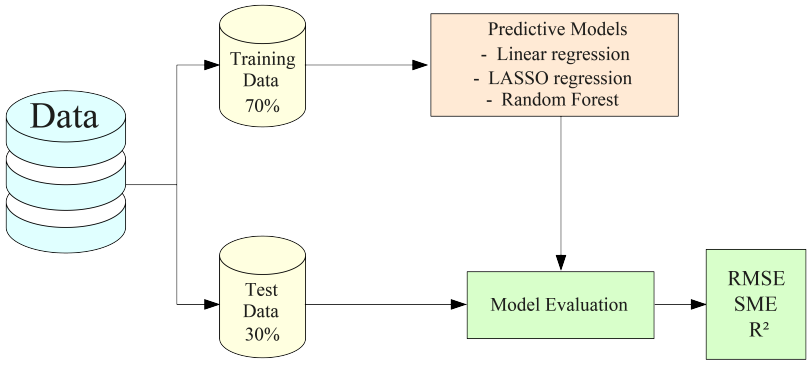
For supervised learning methods, this paper applied three models i.e., multiple linear regression, lasso regression and random forest regression. Firstly, features from the data (consumer comments) were summarized into distinct variables i.e.: unigrams(words), bigrams (two words), LDA topics, sentiments, emotions, categorical variables for influencer categories, dummy variables for emoji, hashtag and handle (@). The data was then split into a training (70%) and a test set (30%).

Thereafter, the features were applied to a linear regression model to predict engagement. Subsequently, the significancy of the coefficients was assessed. To determine variable importance, the t-values were considered. As an alternative, standardised coefficients can be considered, however words that are infrequent, have a low standard deviation and a higher misleading coefficient.

Given that there are many features in the data, it was crucial to mitigate overfitting and also to eliminate noise from the data, so as to attain a more sparse model. Thus, Lasso regression was applied to obtain variables that are important and eliminate features that are not important. Moreover, Lasso also applies a penalty term to the coefficients to minimize variance and therefore enhance prediction.

Random forest was also applied in this research, because tree-based methods typically perform better, when there are many features/variables in the data. Moreover, random forest incorporates interactions and also provides the most important variables in the data. Finally, the models were evaluated based on: R^2 and the Root Mean Square Error (RMSE), using the training and test data sets.

3.3.1 Predictive Model Framework



3.3.2 Multiple Linear Regression

In the multiple linear regression model, the aim is to determine the relationship between features (topics, unigrams, bigrams, etc) and the dependent variable (engagement). Holding other predictors constant, an individual feature's magnitude and significance is then ascertained. Categorical

variables such as the influencer categories and LDA topics, are considered relative to the reference variable.

$$Y = \beta_0 + \beta_i X_i + \beta_j X_j + \beta_k X_k + \cdots + \epsilon$$

where Y is the dependent variable engagement , β_0 is the intercept, β_i is a predictor from the subset of features e.g. unigrams, β_j is also from among the subsets of other features e.g. bigrams, β_k is e.g. a predictor from the categorical features, e.g topics and ϵ is the error term. Interactions are also included in the model, however not with every variable, given the large number of features. Hence, the motivation to apply random forest.

3.3.3 Random Forest Regression

Random Forest is a tree-based method that is an extension of bagging. The method obtains bootstrapped samples from the raw data and grows another tree. In the process of a tree split, independent samples of predictors are taken from the bootstrapped sample. In final aggregate of trees, the model attains decorrelated predictors in the bootstrapped samples and therefore minimizes variance.

$$\hat{f}_{\text{rf}}^{B} = \frac{1}{B} \sum_{b=1}^{B} T(x; \Theta_b)$$

(Hastie, Tibshirani, Friedman, 2017, p.589)

where B is the number of bootstrapped samples, T is the grown random forest tree , x is the point at which a prediction is made , Θ_b is the random forest tree that is split at a node and terminal.

3.3.4 Lasso Regression

Lasso regression shrinks some coefficients to zero and therefore selects variables that are important in an automated process. Lasso is an extension of linear regression, in that, it also incorporates least squares to minimize variance and therefore attain better prediction. This paper applied Lasso to curb the possibility of overfitting, given the large number of features inherent in text data.

$$\begin{aligned} \hat{\beta}^{\text{lasso}} = \quad & \text{argmin}_\beta \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \\ & \text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t \end{aligned}$$

(Hastie, Tibshirani, Friedman, 2017, p.87)

where $\hat{\beta}^{\text{lasso}}$ is the attained Lasso coefficient, y_i is the dependent variable, β_0 is the intercept, x_{ij} are the features in the data , β_j are the coefficients, and t is the parameter for regularisation. Cross

validation is applied to determine the tuning parameter that is optimal to shrink some coefficients to zero and thereafter obtain the important variables.

3.3.5 Predictive Model Evaluation

The R^2 is utilised in this research to evaluate the performance of the regression models, however given the large number of features in the data, it is cautiously considered. Hence, combined with the other metrics(RMSE and MAE), gives a better evaluation framework.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

where R^2 is sum of the squared difference between the dependent variable (engagement) and the predicted variables, ideally R^2 close to 1 would indicate a good fit, y_i is the ith value of engagement , $\hat{y}_i$ is the model predicted value of engagement, $\bar{y}$ is the mean value of engagement.

The Root Mean Square (RMSE) is also used to evaluate the predictive models applied in this research, models are assessed based on the training and test samples.

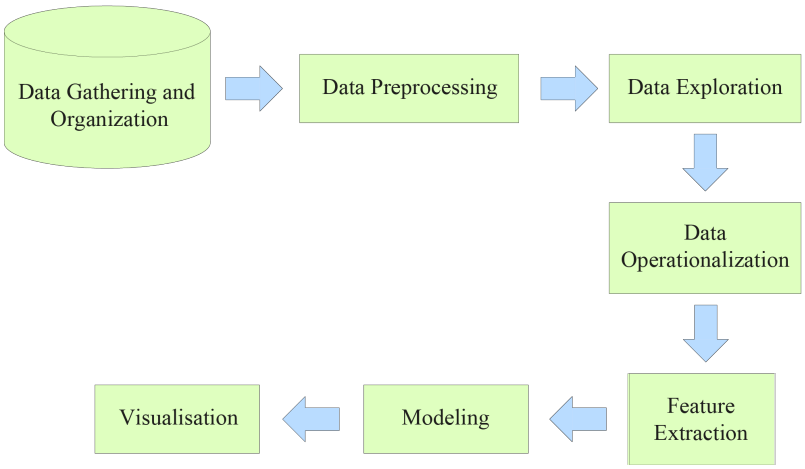
$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

However, Random Forest (RF) is also evaluated on the Out of Bag (OOB) error, given that RF (and other tree-based methods) typically perform better on datasets with a large number of predictors. The model results will be presented and discussed further in chapter 5.

4 DATA ANALYSIS

This chapter elaborates on the data analysis process applied in this research. Starting with the data source, data description, data cleaning, data exploration and finally extracting features for both unsupervised and supervised models. This process is also supplemented with a visualization to illustrate the steps taken to obtain the final models. Figure 4.1 summarizes the process systematically, and each step will be further elaborated separately. This chapter is divided into five sections, each section sequentially explains a particular step taken in the visualisation and the results obtained.

Figure 4.1: Data Analysis Process



4.1 Data Gathering and Organization

The data was obtained from this research paper: `https://dl.acm.org/doi/pdf/10.1145/3366423.3380052/`, where the researchers provide data for further research on Influencer marketing on Instagram. The link to the dataset is here : `https://sites.google.com/site/sbkimcv/dataset/`. The data was crawled from Instagram in a three-month period, during that time, the authors periodically queried Instagram (ad#) to download additional meta data, as influencers updated their posts. Firstly, the Instagram posts were identified as potential Influencer posts based on personal profiles with 1000 or more followers. This (1000) is the generally accepted number to be considered an Influencer (De Vierman et al.2017, Shopify 2021). Secondly, during a brand promotion campaign, influencers are legally required to explicitly indicate with a hashtag i.e. ad# on their posts, that clarify the posts as paid advertisement (US Commission 2017). Thirdly, the researchers only considered influencers with 300 or more posts, during the time of obtaining the data. Therefore, it is assumed that the data is from brand-sponsored influencer marketing campaigns and that the influencer posts contained were for a commercial purpose. The original data provided are nested JSON files. After a series of data wrangling, 1 million (1,000,000) observations were obtained and organized in a data frame. Table 1 summarizes the variables contained in the dataset.

Table 1: Variables from Raw Data

Variable	Description
UserID	Influencer's user ID on Instagram
Username	Influencer's name on Instagram
Comments	Number of comments per influencer's post
Likes	Number of likes on an influencer's post
Category	Influencer's industry e.g., Food, Fashion, Fitness etc
Number of followers	Registered number of influencer's followers on Instagram
Number of posts	Number of posts during the campaign

The User ID is the unique number for each user (influencer and follower) of Instagram. Username is the name indicated on the influencers' and followers' profile accounts. Comments are made on the influencer's post in response to the post's content or other consumers'(followers') comments. Category is the influencer's industry, e.g. Fashion, Food, Travel, etc. From the number of followers, the generally accepted influencer categories on social media, are obtained i.e., Nano, Micro, Macro and Mega influencers (Campbell and Farrell 2020, Haelein et al.2020). Finally, posts are the total number of posts (300 or more) on influencers' profile accounts , during the time data was gathered.

4.2 Data Pre-processing

The data pre-processing involved the following steps: Firstly, this research focused on 3 influencer categories, and excluded Influencers with more than 1 million followers, because that category includes celebrities. Secondly, the data obtained is from influencers and followers from around the world, therefore comments were in multiple languages, only comments in English were retained. Thirdly, 5 (out of 9) categories were selected for this study. There after, meaningless words , numbers and other punctuation were also removed , emojis, hashtags and handles (@), were retained. Words occurring less than 50 number of times were removed. Based on the distribution of words, comments with 37 or more words were retained. Then stop words, and highly frequent words, were removed, given that in topic modelling, such words would appear in all topics. Finally, for the linear regression, a dummy variable was created for emojis, hashtags, and handles (@) and a categorical variable was created for the different influencer groups (Nano, Micro and Macro).

4.3 Data Exploration

After cleaning the data, the next step was to explore the data, to ascertain the distribution of words and the data adequacy for the subsequent unsupervised and supervised learning models. From Table 2, the total number of unique words is 5,392 and the minimum number of words per comment is 51. The word range is sufficient for LDA topic modelling, in that, a wide distribution of

Table 2: Data Exploration

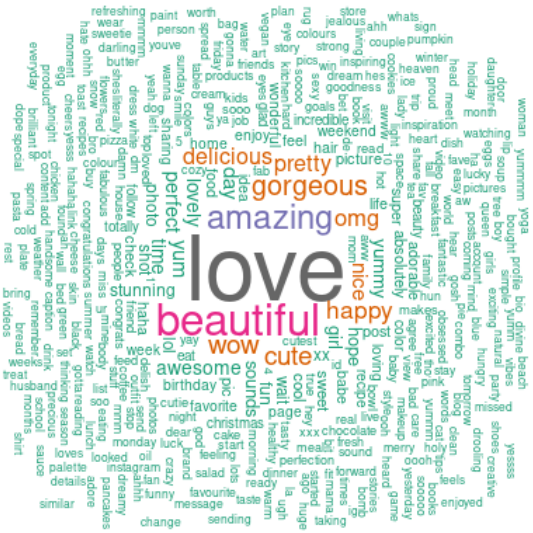
VARIABLE	DESCRIPTIVE STATISTICS	NUMBERS
WORDS	Total number	5392
WORDS PER POST	Minimum	51
	Mean	294.8
	Maximum	13528
WORD OCCURRENCE	Minimum	1
	Mean	25.1
	Maximum	139203

words is required to detect clear topics (Blei et al.2003). For the Glove Model word embeddings, a minimum of 20 words per document (i.e. per comment) is generally considered sufficient to obtain an adequate local context from which the meaning of words can be established (Pennington et al.2014).

The average number of word occurrence is 25 and the maximum is 139,203. This extensive range is necessary to extract: unigrams, bigrams and sentiments (polarity), as features that will be applied in predictive models. An adequate variation of text data enables one to determine the relationship between the dependent variable (engagement) and the subset of features e.g., sentiments, emotions. Thus, in the final analysis, the particular drivers of engagement can be established.

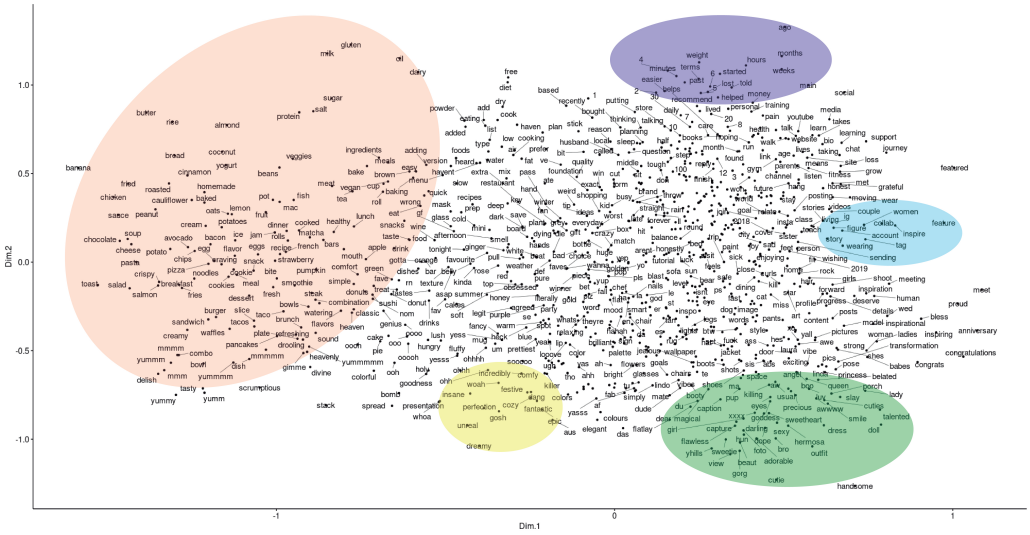
To explore the data further and have a better understanding of the data, two visualizations are presented below, a word cloud (Figure 4.2) that contains 400 most frequent words in the data and Figure 4.3, a Multi-Dimensional Scaling (MDS) map of word co-occurrences (400 most co-occurring words).

Figure 4.2: Word cloud



The word cloud shows that love, beautiful and amazing are the most frequent words in the data, and next to that, are words like: wow, cute, gorgeous, pretty, girl, photo, picture, etc , suggesting that the words are about a female person in a photo, perhaps advertising or showcasing something. There are also words like: delicious, yummy, recipe, food, etc, indicating food related terms. However, zooming in on the word cloud, there are words with scattered meaning e.g home, paint, space, hair, dress and makeup, etc, words that are about home interior, and fashion/cosmetics i.e. indicating scattered themes. MDS minimises the distance between words, such that similar words appear next to each other or latent themes within the data are revealed. The MDS word map below indicates distinguishable themes with in the data.

Figure 4.3: MDS map of word co-occurrences



From the MDS map, on the left (peach color theme), the words are all clearly about food, e.g. banana, veggies, milk, etc., and to the bottom right (green color theme), the words are about women's outlook/ fashion, e.g., dress, outfit, shoes, beaut, sexy, etc. The clusters can be better understood, by considering the words within and around the illustrators. In the north west (dark blue) there are words like: weights, hours, days , etc, and around the cluster: personal training, gym, pain, etc, suggesting that the topic of discussion is about personal training/body fitness. Thus, from the MDS visualization, themes with in the data are established, this then requires the next step to build a more robust model to detect the topics within the whole data i.e. LDA topic modelling.

4.4 Data Operationalization

The dependent variable in this study is engagement i.e., comments and likes on an influencer's post. Engagement is thus consumers' responses to the influencer's content and other consumers' reactions. In the marketing literature e.g., Hughes et al.2019 and Jaakonmäki et al.2017, engagement is operationalized by counting the number of comments and likes and then dividing that by the

number of followers. This controls for the absolute differences among the numbers of influencers' followers. Multiplying by 100%, makes it a comparable measurement metric (engagement rate) to measure and compare engagement across the various influencer categories. This research follows that and operationalizes engagement as indicated below:

$$\text{Consumer engagement rate} = \frac{(\text{ Number of comments } + \text{ Number of likes })}{\text{Number of followers}} * 100\%$$

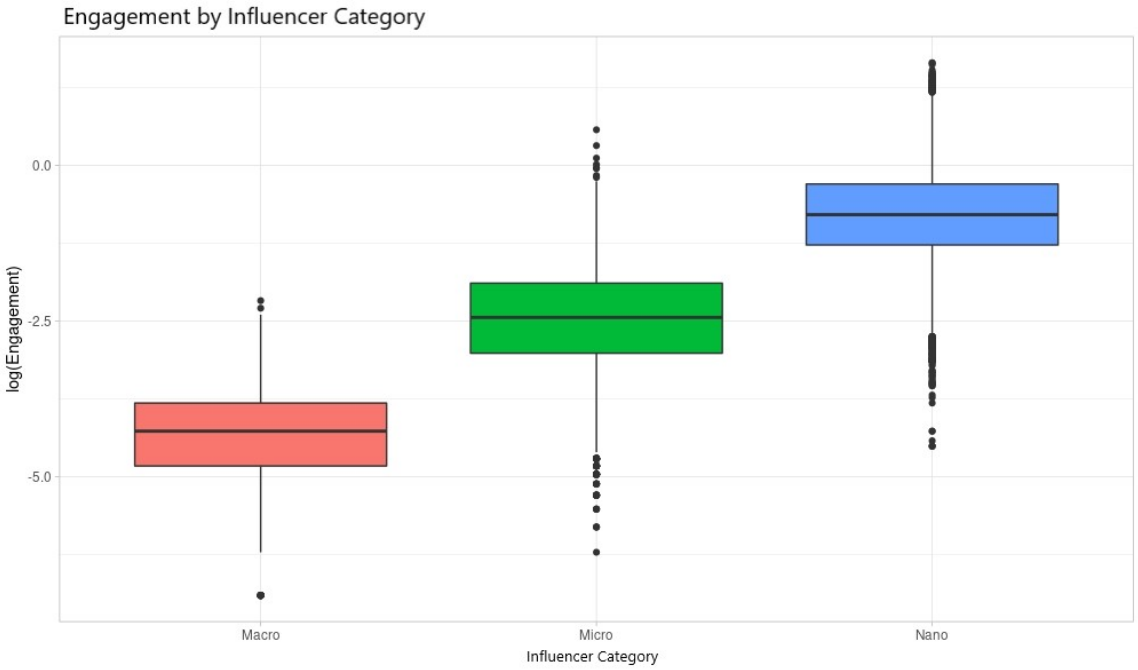
The formula above is applied to each influencer post in the data, and the results are summarized in Table 3 below:

Table 3: Engagement rates

VARIABLE	**DESCRIPTIVE STATISTIC**	**NUMBER**
NUMBER OF LIKES	Minimum	0
	Mean	9.084
	Maximum	419
NUMBER OF COMMENTS	Minimum	1
	Mean	19.47
	Maximum	40
NUMBER OF FOLLOWERS	Minimum	1001
	Mean	28158
	Maximum	940814
ENGAGEMENT RATE MICRO INFLUENCERS	Minimum	0.0021%
	Mean	0.0757%
	Maximum	0.3979%
ENGAGEMENT RATE NANO INFLUENCERS	Minimum	0.0107%
	Mean	0.3874%
	Maximum	3.7249%
ENGAGEMENT RATE MACRO INFLUENCERS	Minimum	0.0005%
	Mean	0.0108%
	Maximum	0.0399%

The number of followers is between 1001 to 940,814, this study excludes Mega influencers (influencers with over 1 million followers), thus the number of comments and likes are relatively modest. Likes are between 0 to 419 with an average of 9 likes. Comments are between 1 and 40, with a mean average of 19. Hence, in this study, there are on average more comments per post than likes (in total absolute terms, likes are more). Nano influencers have the highest engagement rate of 3.7%, followed by Micro (0.4%) and Macro (0.04%). The results are consistent with some studies in the literature e.g., Park et al.(2021) and Kay et al.(2020) compared Micro vs Macro, and their findings are that Micro have a higher engagement rate compared to Macro. Which is also the finding of this

Figure 4.4: Consumer Engagement rate by influencer category



study.

The boxplots are logarithmic transformations of the engagement rate results obtained in Table 3. As can be noticed, Nano influencers have a higher engagement rate compared to Micro and Micro influencers. This finding answers research question (2): "Comparatively, are there differences in engagement rates, between Nano, Micro and Macro Influencers on Instagram?" This research question will be further examined with the predictive models in which the influencer categorical variables are a subset of features in the model.

4.5 Feature Extraction

Features extracted from the data are summarized in Table 4 below. The predictors will be applied in a prediction model as potential drivers of engagement. Firstly, influencer categorical variables are applied in the predictive model to control for the differences in influencer's number of followers, given that this comparably generates more engagement (Bahkshi et al.2014). Secondly, the aim of this study, is to compare engagement rates across influencer categories, through t-tests, the results indicated above (engagement rates), and predictive model results, research question (2) will be thoroughly answered.

Tirunillai and Tellis (2014) investigate online product reviews by using Latent Dirichlet Allocation (LDA). The researchers find that detected consumer topics enable marketers to track key latent dimensions of consumer satisfaction, from which brands can attain competitive positioning. Secondly, Kumar et al.(2013) argue that consumer WOM is a key variable in attaining a successful marketing

Table 4: Gathered Features from Data

Features	**Feature summary description**
Unigrams	50 most frequent words
Bigrams	50 most frequent co-occuring words
LDA topics	15 LDA topics detected from data
Sentiments	Positive, Neutral and Negative sentiments from Polarity Dictionary
Emotions	8 emotions (trust, anger, fear, sadness, surprise, joy, anticipation, disgust) from NRC Library
Number of words	Number of words in the comments
Dummy variables	Dummy variables for emojis, hashtags, handle (@)
Influencer Categories	Categorical variables for the 3 influencer-groups
Influencer Posts	Number of posts by each influencer

campaign. Therefore, this research applies LDA topics as features to ascertain their significancy in driving consumer engagement based on research in online User Generated Content (UGC).

Jaakonmäki et al. (2017) investigated consumer engagement on Instagram, there findings were that, emojis are significant drivers of engagement. Furthermore, Trusov et al.2009 compared consumer-to-consumer and brand-to-consumer engagement and found consumer-to-consumer to garner significantly higher engagement than brand engagement. Consumer-to-consumer engagement (e.g. in comments online) can be ascertained through the use of @ in consumer comments. Additionally, Suh et al. (2010) investigated retweeting on social media (Twitter), there findings are that the use of hashtags significantly increases the retweet rate. Thus, based on the mentioned studies, emojis, @, hashtags, are gathered and applied in the predictive model as potential drivers of engagement.

Consumer engagement literature (Pansari and Kumar 2017, Santani et al.2020) consider emotions as key factors in fostering WOM online. Emotions such as anger, joy, and trust are known to generate interactions on social media e.g., Berger and Milkman (2012) find that virality on social media is generated by content that evokes awe and anger. The NRC library in R contains 8 emotions (anger, joy, trust, sadness, anticipation, disgust, surprise and fear). Thus, emotions are also examined in a prediction model.

Consumer sentiments i.e., positive and negative opinions are extracted because they can be a rich source of knowledge to brands e.g., in product reviews or in online brand communities. Knowledgeable consumers contain product knowledge and consumer preferences. Such knowledge can be an effective means for product development for firms (Kumar et.2010, Tirunillai and Tellis 2014).

5 RESULTS AND ANALYSIS

5.1 Latent Dirichlet Allocation (LDA) Results

The results of the unsupervised and supervised learning methods applied in this research are discussed in this chapter. Firstly, unsupervised models (LDA , Glove Model, and a comparison of the two models). Secondly, the chosen model and in addition to other features, are all then examined in supervised learning models: Linear regression, Lasso and Random Forest . Finally, an evaluation of the prediction models applied in this study, is presented at the end.

The LDA topics were obtained based on the coherence value, in appendix A, the three ANOVA tests indicate that k = 15 was taken as the optimal number of LDA topics. Figure 5.1 shows the topics detected by LDA displaying the top most frequent words in each topic. The topics are generally coherent and can be interpreted. 5 topics are briefly elaborated below:

The topic “beauty industry” has terms such as: hair, makeup, product, skin, brand, wear, video. This topic suggests a topic discussion of women’s beauty brand products and a video of how to wear the cosmetic.

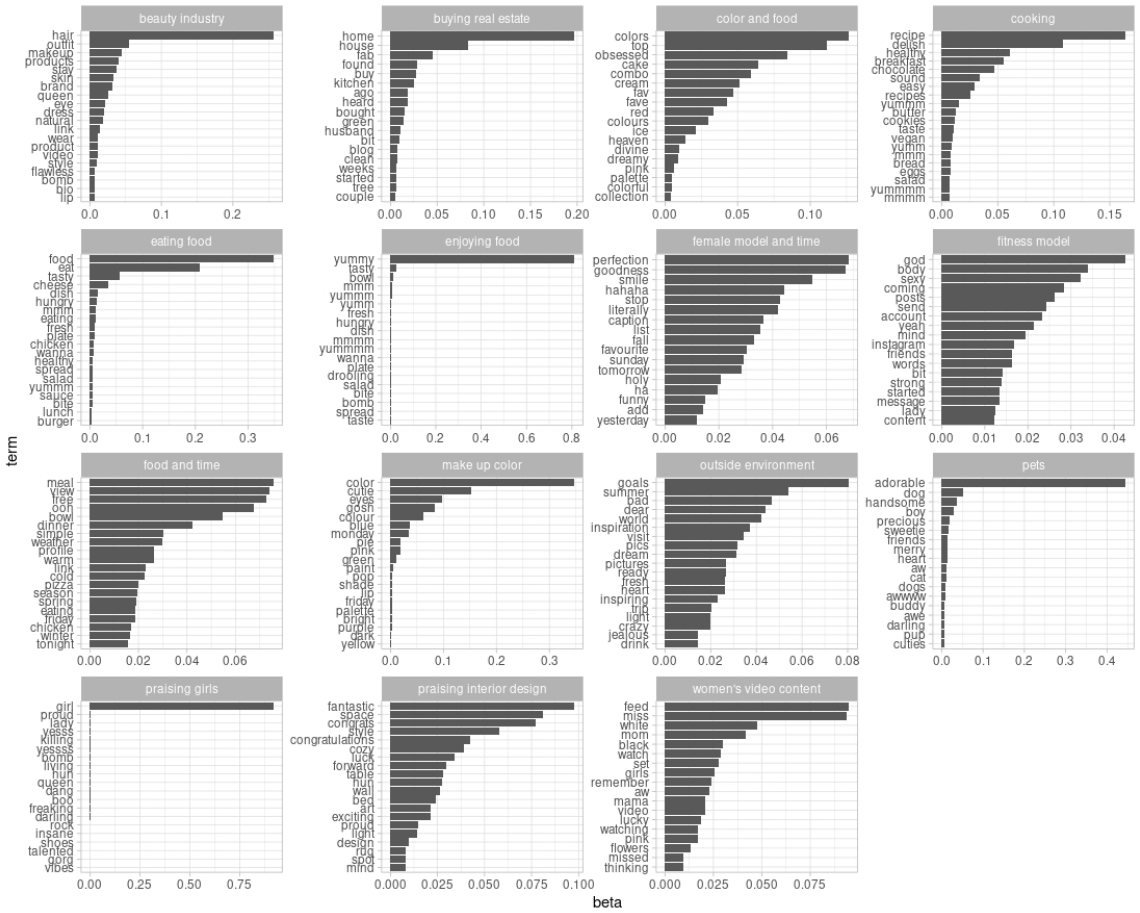
The topic “eating food” has terms: food, eat, tasty, lunch, sauce, chicken, salad, spread, fresh plate, healthy, burger. This topic is about lunch as a meal and the type of food eaten for lunch.

The topic “pets” has terms such as: adorable, dog, cat, dogs, buddy, pup, cuties, darling, precious, heart. This topic is about pets and what they mean to the owners.

The topic “praising interior design” has terms such as: fantastic, space, wall, table, bed, art, style, cozy, luck, design, light, proud, exciting. This topic is about describing the interior design of the house and the facilities admired.

Finally the topic ”cooking” has terms: recipe, delicious, healthy, breakfast,bread, butter, eggs, salads, easy, recipes. This topic suggests a discussion of the recipes to make a healthy breakfast.

Figure 5.1: LDA topics



5.2 Glove Model Results

The Glove Model applied was based on a window of 6 words, before and after the focal word and d = 50 (dimension). The parameters are based on the suggestions of Penning et al.(2014), (the authors of the Glove Model) where the window taken is 10 words. However, without stop words, and depending on the size of the corpus, a smaller window gives better results. The dimensions are usually based on the size of the corpus. Table 5 shows the word embeddings obtained from the data.

The Table shows 200 words selected from the Glove Model co-occurence matrix (word embeddings)

The words are arranged in the table in descending order such that words with the highest word embedding (cosine similarity) appear in the first row of the column.

The table of words indicates that embeddings that are similar equally have words with similar meaning. For example, in the 4th column (dogs) words such as: pup, puppy, cat, dogs, are clearly about pet animals. While in the 3rd column (hair) words such as: make up, skin, curls, body, etc,

are all about the beauty industry and associated terms. In the 6th column (home) words: house, kitchen, space, room, are all about a home or facilities in a home. The same pattern can be noticed in the other columns. However, in each column, there are words that are specific to the context e.g., in the 1st column, words such as motivation, business, Instagram, etc are all general words but also peculiar to the fitness industry. In the 2nd column (girl) words such as omg (ooh my god), damn, wow, etc,are words generally used in everyday speech to express an impression, but are also specific to this context, in that, they seem to be a response to a picture displayed. The same pattern can also be noticed across other columns, i.e. Without the context, the words can seem incoherent to the others.

Table 5: Glove Embeddings

fitness	**girl**	**hair**	**dog**	**food**	**home**	**colour**	**outfit**
health	babe	makeup	pup	healthy	house	color	dress
0.6683	0.8157	0.791	0.7991	0.7119	0.8447	0.8988	0.8185
workout	mama	skin	cat	yummy	kitchen	shade	hat
0.5706	0.7998	0.7541	0.7727	0.6362	0.7667	0.7489	0.7286
training	sis	curls	puppy	meal	room	pink	bag
0.5693	0.6873	0.6963	0.7313	0.6258	0.7511	0.704	0.7142
instagram	omg	style	dogs	yum	space	lipstick	suit
0.5636	0.6872	0.6257	0.7107	0.6243	0.6697	0.6764	0.7118
business	lady	body	cutie	stuff	garden	palette	shirt
0.5612	0.6856	0.6111	0.6864	0.6113	0.6524	0.6742	0.6703
motivation	beauty	natural	boy	delicious	tree	blue	jacket
0.5565	0.678	0.6051	0.6821	0.5927	0.6439	0.6655	0.6585
yoga	damn	color	baby	foods	family	texture	shoes
0.5555	0.672	0.5978	0.6588	0.5897	0.6408	0.6647	0.6576
page	boo	beauty	guy	looks	bed	paint	style
0.5399	0.6482	0.568	0.6552	0.5852	0.6235	0.6625	0.6525
diet	wow	flawless	little	time	bedroom	colors	babe
0.5248	0.6324	0.5628	0.6499	0.5846	0.6128	0.6374	0.6445
follow	lol	lips	awww	way	christmas	yellow	gorg
0.5093	0.6262	0.5604	0.6458	0.5817	0.603	0.6295	0.6403
ig	seriously	eye	face	eating	place	rug	pants
0.4915	0.6209	0.5516	0.6293	0.5803	0.6022	0.6207	0.6356
body	gorgeous	lip	mom	best	back	colours	caption
0.4797	0.6144	0.5479	0.6204	0.5784	0.5814	0.6137	0.5841
travel	beautiful	sis	handsome	recipes	time	light	slay
0.4721	0.6044	0.5424	0.6062	0.5772	0.578	0.6091	0.5828
group	absolutely	queen	aww	salad	cozy	print	pic
0.4662	0.6014	0.5404	0.5921	0.5693	0.5573	0.605	0.573
journey	queen	face	adorable	plate	table	wood	sweater
0.4503	0.5954	0.5323	0.5835	0.5683	0.5536	0.5663	0.5669
gym	pic	nails	precious	spread	style	wall	girl
0.4498	0.586	0.5267	0.5757	0.5671	0.5415	0.5629	0.551
weight	freaking	outfit	lil	comfort	city	stunning	cute
0.4469	0.5825	0.5267	0.5705	0.5664	0.5346	0.5513	0.5509
profile	amazing	cut	picture	good	area	door	shot
0.4459	0.578	0.5262	0.5606	0.5603	0.5331	0.5449	0.5507
content	pretty	lipstick	cute	cooking	decor	gorgeous	queen
0.4436	0.5737	0.5228	0.5588	0.5594	0.5314	0.5399	0.5427
support	awesome	eyes	sweet	fruit	husband	pretty	luv
0.4366	0.5706	0.5146	0.5492	0.5552	0.5254	0.5347	0.5403
shape	hun	foundation	smile	right	enjoy	walls	makeup
0.4309	0.5649	0.5144	0.5351	0.553	0.5187	0.5317	0.5339
check	baby	lashes	awwww	breakfast	clean	mirror	mug
0.4271	0.5624	0.5118	0.5245	0.5529	0.5106	0.5256	0.5296
vegan	shot	dress	lol	feed	move	grey	pretty
0.4227	0.5602	0.5103	0.5146	0.5451	0.5014	0.5249	0.5277
strong	goals	slay	friend	really	year	dress	hair
0.4145	0.5566	0.5036	0.5128	0.5448	0.5011	0.5238	0.5267
form	guys	girl	name	like	friend	red	pose
0.4141	0.5562	0.5029	0.5053	0.5423	0.4983	0.5201	0.5264

5.3 LDA and Glove Model

LDA takes a global approach and therefore topics detected are broad themes in the corpus without the local context. On the hand, the Glove Model considers the semantic and syntactic properties of the words, such that word similarity is derived from words that occur together (syntactic or local context) but also words that are similar (semantic). The word similarity is based on the cosine similarity. K-Means was applied to the Glove Model co-occurrence matrix to ascertain whether the clusters obtained are intuitive and similar to the Glove model word embeddngs matrix. There after compare the K-Means clusters with LDA topics. To compare LDA and the Glove Model, K-Means was applied with same the same k number of clusters as the LDA topics, i.e. k= 15. The aim is to determine the unsupervised model to apply in the prediction models.

Table 6 shows the K-Means labeled Clusters and their associated words.

Table 6 shows that the clusters obtained are in some cases interpretable while in other cases a clear label could not be established. The intuitive clusters are e.g. 6 “Beauty Industry” that has similar terms with “Beauty industry” in LDA and Cluster 3 “Pets and Social life” with somewhat similar terms to “pets” in LDA. However, the other clusters, no clear coherence can be determined.

Firtstly, the difference between the Models lies in the level of details in the Glove Model clusters, e.g. in “ praising” the terms are in reference to an object in a way that they (terms) refer to what is noticed or happening with some detail, while in LDA e.g. in “praising girls” there are terms : girls, lady, etc , i.e. the terms revolve around the object (girls, lady). Also in “ pets and social life” for the Glove cluster, there are terms like daily, every day, perhaps indicating what is done i.e. walking the dogs, while in “pets” in LDA, that level of detail is absent.

Secondly, in the Glove Model, there are no high frequency occurring words, but rather context specific words. For example in LDA topics: eating food, enjoying food, and cooking, there are high frequency words like yummy, taste(y) which occur in a number of topics, while in the food-related Glove Model clusters: cooking methods and food ingredients, such high frequency words (yummy, taste) are not among the terms, but rather the terms are context specific. This can be attributed to the Glove Model’s control for high frequency terms, i.e. the least squares function.

Thirdly, interpretation of the Glove Model is difficult to pinpoint to a clear theme or topic, compared to LDA topics. In LDA topics, broad themes can be established, for example in topics such as eating food, pets, cooking, praising interior design, etc., it is clear what the object of discussion is, whereas in Glove Model clusters, the nuances and details involved are descriptive without pinpointing an object. For example the LDA topic cooking is about recipes for a meal i.e. breakfast, while in the Glove Model cluster: recipes, the terms are not clear, for what meal the recipes are for.

In sum, LDA produces more interpretable topics, while the Glove Model is advantageous, in that, it is context specific, has more fine details, and controls better high frequency terms, hence less noise.

However, because of the difficulty in interpretation, LDA topics are chosen for further analysis (predictive modelling).

Table 6: K-Means Glove Model Clusters

K-Means Clusters for Glove Models Word Embeddings

	Cluster Label	Cluster terms
1	"Body"	• Body related terms *abs, legs, eyes, curls, lips, nails, feet, lashes.* • Make up related terms *shades, yhills*
2	"Internet"	• Internet related terms *reply, comment, tag, click, http, tutorial, review* • Miscellaneous related terms *form, terms*
3	"Pets and Social Life"	• Pet related terms *dogs, cats* • Social Life related terms *friends, daily, everyday, boys* • Miscellaneous related terms *entire, reminds, kinda favorite.*
4	"Verbs 1"	• Verbs related terms *waiting, planning, hoping, buying, starting, missing, posting, coming*
5	"Verbs 2"	• Verbs related terms *growing, telling, call, told, called, grow.*
6	"Beauty Industry"	• Beauty Industry related terms *skin, makeup, hair, product, mask, foundation, natural, brand, line*
7	"Commerce"	• Commerce related terms *store, shop, local, app, restaurant, market, company, quality* • Internet related terms online, app
8	"Fitness Model"	• Fitness related terms *strong, fit* • Model related terms *princess, goodness, angel, model, doll* • Miscellaneous related terms *golden, pure, human*
9	"Miscellaneous 1"	• Clothing related terms *clothes, wear* • Miscellaneous terms *match, team, hold, throw, stand, stick, pull, cut*
10	"Cooking Methods"	• Food description terms *baked, fried, crispy* • Food related terms *waffles, fries, pancakes, noodles, potatoes, bacon*
11	"Motivational Fitness"	• Fitness related terms *struggle, goals* • Motivational related terms *quote, caption* • Agreement terms *understand, agreed, amen* • Miscellaneous terms *truth, funny*
12	"Food Ingredients"	• Ingredient related terms *salt, sugar, milk, oil, butter, almond, lemon, coconut, chocolate peanut.*
13	"Recipes"	• Recipe related terms *clean, simple, healthy, fresh, cooked* • Food related terms *fruit, meat, steak, food, oats*
14	"Praising"	• Praising related terms *epic, impressive, dope, legit, insane, unreal, killer, whoa, bomb, unreal*
15	"Miscellaneous 2"	• Intelligence related terms *gift, genius, brilliant* • Catalogue terms *menu, list.* • Miscellaneous terms *project, challenge, level* • Time related terms *tonight, tomorrow*

5.4 Linear Models

Given the relatively high number of features inherent in text data, the features extracted from the data, were first applied to the linear regression model, individually,so to ascertain whether as a subset of features, they could lead to over fitting. Secondly, by considering them individually, the potential relationship of a particular predictor to the dependent variable can be systematically analysed.

Table 7 shows the summarized features applied in the model.

From the table, word interactions have the lowest AIC of 24460, followed by the model with only with unigrams (24833). Perhaps what is some what more informative is that topics and emotions have a relatively high AIC (26144 and 26071) compared to all other features. The consumer engagement literature e.g. Santini et al.(2020) and Pansari Vivek et al.2012, emotions are considered a key factor in generating word of mouth (WOM) and therefore consumer engagement. Further examination of features in the prediction models will clarify the variables that are significant drivers of engagement.

Table 7: Linear Models of Subsets of Features

Model	**df**	**AIC**
Linear model all variables	130	2591.806
Linear model no emotions	112	2592.983
Linear model only emotions	11	26071.517
Linear model only unigrams	50	24833.141
Linear model only bigrams	52	25828.482
Linear model unigrams and bigrams	100	24781.040
Linear model only topics	16	26144.863
Linear model word interactions	1228	24460.68

The above subsets of features are only informative to a certain extent, thus in order to establish the relationship of all features to the dependent variable, all predictors were applied to a multiple linear regression model. Secondly, the influencer categories were subjected to t-tests i.e. Micro vs Nano, Macro vs Nano and Micro vs Macro influencers, the results (in Appendix B) indicate that, in this study, Nano influencers have a higher engagement rate compared to Micro and Macro. However, by applying the influencer categories in the prediction model, a clear finding can be established.

The Linear Model is presented in Table 8, with all the features extracted from the data. The significance of the features can be interpreted while holding all other variables constant. Thus, a feature is considered as a driver of engagement, without the interactions or the effect of the other features.

From the table, Nano influencers have a relatively higher engagement, compared to the reference category, i.e. Macro influencers, This result is consistent with earlier results, i.e. t-tests (Appendix B) , engagement rates and boxplots in chapter 4. Thus, the combination of the aforementioned results, answers research question (2) " comparatively are there differences in engagement between the influencer categories"? This finding is also inline with Park et al.(2021) and Kay et al.(2020) who find that Micro have a higher engagement compared to Macro influencers. The results will also be compared to the other prediction models.

Emojis are in this research, also significant with a positive coefficient. This finding is also in line with literature on consumer engagement, e.g. Jaakonmaki et al.(2017) in a study of engagement on Instagram, also find emojis to be significant drivers of engagement. Given that Instagram is a low-involvement media (vs high involvement e.g Blogs), features such as emojis may interact with the type of media, hence models with interactions delineate that effect.

Handles (@) are also significant in this study, with a positive and significant coefficient. The significance of handles is also intuitive, in that , on social media, consumers interact with each other and therefore propagate the WOM , hence driving engagement further. Consumer engagement literature posits that consumer-to-consumer engagement generates higher engagement as compared to brand-to-consumer (Trusov et al.2009), thus this study aligns with that finding.

However, number of posits, have a negative and significant coefficient, hence a negative effect on engagement. Given that the data was gathered over a three-month period, the possible explanation can be that , over time consumers became fatigued by the influencer posts. This is the finding of Belanche et al.(2017) and Bakalash and Reimer 2013, who argue that visual images engage consumers to a certain extent and thereafter diminishing returns take effect. Perhaps, alternatively quality vs quantity prevails.

Topics 5 and 14 ("eating food" and "praising interior design" respectively) have a positive effect on engagement compared to the reference category, i.e. topic 15. This result is also in line with literature , but also the media type i.e. Instagram is a visual image media. Food is one of the most popular industry on Instagram and the two overarching concepts in the food industry are arousal and stimuli, and according to research, both are fostered by visual images (Holmqvist and Lunardo 2015).

Table 8: Linear Model

term	estimate	std.error	statistic	p.value
(Intercept)	-0.0490902	0.0095639	-5.1328586	0.0000003***
Number of words	0.0002061	0.0000238	8.6628050	0.0000000***
"Emoji"yes	0.0210736	0.0027746	7.5952902	0.0000000***
"Hashtag"yes	-0.0010842	0.0026793	-0.4046568	0.6857319
"Handle"yes	0.0122525	0.0027286	4.4904750	0.0000071***
Micro Influencers	0.0846641	0.0035550	23.8155561	0.0000000***
Nano Influencers	0.5908653	0.0044068	134.0805418	0.0000000***
Number of posts	-0.0000175	0.0000010	-18.0795982	0.0000000***
love	0.0073088	0.0009229	7.9193597	0.0000000***
amazing	0.0088275	0.0014811	5.9599689	0.0000000***
wow	0.0112294	0.0037134	3.0240101	0.0024960***
cute	0.0054986	0.0012570	4.3744446	0.0000122***
pretty	0.0095134	0.0018611	5.1116590	0.0000003***
delicious	0.0062494	0.0018460	3.3854600	0.0007113***
nice	0.0145674	0.0028077	5.1883323	0.0000002***
yum	0.0169976	0.0032814	5.1799308	0.0000002***
awesome	0.0120118	0.0027868	4.3102728	0.0000163***
yummy	-0.0136661	0.0045680	-2.9916689	0.0027762***
girl	0.0093954	0.0021187	4.4345719	0.0000092***
stunning	0.0108772	0.0034631	3.1408584	0.0016857***
shot	0.0163345	0.0028275	5.7769039	0.0000000***
xx	0.0047437	0.0014617	3.2453601	0.0011740***
cool	0.0124691	0.0027796	4.4859495	0.0000073***
pic	0.0044774	0.0021884	2.0459259	0.0407701**
food	0.0053577	0.0025684	2.0860004	0.0369848**
feel	-0.0045090	0.0022135	-2.0370426	0.0416523**
babe	0.0199089	0.0028487	6.9887124	0.0000000***
enjoy	-0.0106489	0.0038770	-2.7467211	0.0060221***
gorgeous.love	0.0632623	0.0283698	2.2299139	0.0257586**
pretty.love	0.1328734	0.0374802	3.5451620	0.0003928***
amazing.wow	0.1370306	0.0513560	2.6682484	0.0076278***
nice.pic	-0.0389122	0.0160456	-2.4251052	0.0153083**
ice.cream	0.0148698	0.0052116	2.8531986	0.0043304***
yum.love	0.1768799	0.0723146	2.4459775	0.0144502**
nice.shot	0.0941957	0.0201362	4.6779297	0.0000029***
gorgeous.shot	-0.0414761	0.0190331	-2.1791588	0.0293256**
girl.love	0.2262292	0.0443935	5.0960037	0.0000003***
Polarity sentiment score	0.0105554	0.0078216	1.3495187	0.1771780
Anger emotion score	0.3369289	0.1502271	2.2427963	0.0249153**
Anticipation emotion score	0.0757425	0.0768628	0.9854243	0.3244216
Disgust emotion score	0.0298451	0.1537414	0.1941251	0.8460789
Fear emotion score	0.0071595	0.1462792	0.0489443	0.9609639
Joy emotion score	0.0753013	0.0442071	1.7033750	0.0885056*
Sadness emotion score	-0.2443479	0.1265756	-1.9304502	0.0535581*
Surprise emotion score	0.0374482	0.0903355	0.4145455	0.6784769
Trust emotion score	-0.0641482	0.0780304	-0.8220927	0.4110290
Topic 1	-0.0277939	0.0106049	-2.6208569	0.0087742***
Topic 2	-0.0072730	0.0092073	-0.7899123	0.4295836
Topic 3	-0.0255396	0.0105684	-2.4166133	0.0156700**
Topic 4	0.0043131	0.0139333	0.3095534	0.7569022
Topic 5	0.0366370	0.0103572	3.5373552	0.0004045***
Topic 6	-0.0228740	0.0138098	-1.6563639	0.0976559*
Topic 7	0.0425037	0.0095291	4.4604370	0.0000082***
Topic 8	-0.0123037	0.0126836	-0.9700441	0.3320303
Topic 9	-0.0171461	0.0080090	-2.1408625	0.0322893**
Topic 10	-0.0229722	0.0082467	-2.7856345	0.0053440***
Topic 11	-0.0129484	0.0113163	-1.1442239	0.2525376
Topic 12	-0.0252575	0.0080977	-3.1191002	0.0018149***
Topic 13	-0.0168226	0.0098561	-1.7068195	0.0878633*
Topic 14	0.0343477	0.0113133	3.0360401	0.0023981***
Observations	58,012			
R^2	0.4435			
Adjusted R^2	0.4418			
Residual Std. Error	0.2494 (df = 40488)			
F Statistic	271.1***(df = 119; 40488)			
				Note: p<0.1*; p<0.05**; p<0.01***

Table 9: Linear Model with Interactions

term	estimate	std.error	statistic	p.value
(Intercept)	-0.0103329	0.0400011	-0.2583161	0.7961644
Nano Influencers	0.4373425	0.0275799	15.8573001	0.0000000***
Nano Influencers x love	0.0379917	0.0030921	12.2867679	0.0000000***
Nano Influencers x perfect	0.0655978	0.0079022	8.3012459	0.0000000***
Nano Influencers x omg	0.0962485	0.0146113	6.5872751	0.0000000***
Nano Influencers x beautiful	0.0314845	0.0048433	6.5005940	0.0000000***
Nano Influencers x lol	0.0712900	0.0115311	6.1824012	0.0000000***
“Emoji”yes x yum.love	7.3569848	1.3037761	5.6428282	0.0000000***
Nano Influencers x nice	0.0547257	0.0097563	5.6092556	0.0000000***
ice.cream x Topic_4	0.3201316	0.0599962	5.3358613	0.0000001***
Number of words x joy emotion score	0.0054706	0.0010333	5.2945227	0.0000001***
Nano Influencers x pretty.love	1.4160304	0.2750893	5.1475299	0.0000003***
girl.love x Topic 14	-8.5742088	1.6786700	-5.1077394	0.0000003***
Anticipation emotion score x yum.love	96.3890479	19.1220697	5.0407225	0.0000005***
Disgust emotion score x yum.love	-401.6307349	80.4776400	-4.9905879	0.0000006***
girl.love x Topic 10	-7.4639579	1.4964480	-4.9877830	0.0000006***
beautiful.pic x Topic 10	1.1317307	0.2279331	4.9651870	0.0000007***
“Hashtag “yes x yum.love	8.2313082	1.6636086	4.9478636	0.0000008***
Number of words x yum.love	0.1105381	0.0223651	4.9424322	0.0000008***
Nano Influencers x nice.shot	0.4136031	0.0841088	4.9174771	0.0000009***
Nano Influencers x amazing	0.0259330	0.0053189	4.8755914	0.0000011***
yum.love	-63.5204277	13.0384550	-4.8717757	0.0000011***
Nano Influencers x yum.love	30.6458030	6.3210574	4.8482083	0.0000013***
Nano Influencers x birthday	0.0559879	0.0115960	4.8282009	0.0000014***
Micro Influencers x yum.love	22.4874113	4.7049919	4.7794792	0.0000018***
girl.love x Topic 7	-3.9640888	0.8453459	-4.6893097	0.0000028***
Nano Influencers x Number of posts	0.0360867	0.0077349	4.6654204	0.0000031***
Surprise x nice.shot	-8.2788702	1.7770507	-4.6587699	0.0000032***
Anger emotion score x yum.love	682.7209627	147.9156831	4.6156090	0.0000039***
Number of posts x yum.love	0.0061668	0.0013777	4.4761281	0.0000076***
Sadness emotion score x nice.shot	9.9078319	2.2963546	4.3145914	0.0000160***
Fear emotion score x beautiful.photo	7.9150669	1.8587773	4.2582116	0.0000207***
Nano Influencers x girl	0.0309760	0.0073333	4.2240010	0.0000241***
girl.love x Topic 2	-2.5724607	0.6129134	-4.1971032	0.0000271***
Nano Influencers x babe	0.0450460	0.0107351	4.1961483	0.0000272***
Nano Influencers x shot	0.0469586	0.0112808	4.1627201	0.0000315***
Nano Influencers x favorite	0.0453125	0.0109760	4.1283289	0.0000366***
Nano Influencers x Topic 3	-0.1450613	0.0357949	-4.0525726	0.0000508***
Observations	58,012			
R2	0.5412			
Adjusted R2	0.5113			
Residual Std. Error	0.2334 (df = 38124)			
F Statistic	18.11***(df = 2483; 38124)			
				Note: p<0.1*; p<0.05**; p<0.01***

Table 9 is a model with interactions, again Nano influencers interact with many variables positively, e.g with bigrams such as beautiful,babe, short,etc. Perhaps in contrast to the linear model, hashtags interact positively and therefor have a positive effect on engagement. This also is inline with findings from other studies e.g. Suh et al.2010, who also find hashtags and URLs to be key drivers of retweets on Twitter. Hashtags are utilized actively on Instagram for brand and other campaigns.

However, given the number of features in the data, the interactions in the linear model, do not indicate the most important variables and whether there is overfitting, hence the next model applied in this study is Lasso regression.

5.5 Lasso Regression

Lasso regression was applied in this study, firstly by determining the optimal lambda to shrink some coefficients to zero and therefore attain a more sparse model. The optimal lambda was obtained through cross validation and in Appendix D, various values of lambda are indicated and the coefficients at different lambdas. In line with earlier results, emojis, handles (@), and Nano influencers have positive effects on engagement.

Consistent with the linear model, the number of posts have a negative effect on engagement, hence confirming that the finding that, in this study, more posts have a negative effect on engagement.

Emotions interact with mixed effects, positively i.e.joy and Nano influencers, and more so with anger, while joy (with posts), micro influencers and sadness, the effect is negative. The effect of emotions as a subset of features, however requires a delineation of emotions, Because emotions are categorised as positive and negative, thus the effect of one emotion can be indirectly determined by the total effect of the category of emotion (negative or positive). Thus, in combination with the AIC results of emotions,

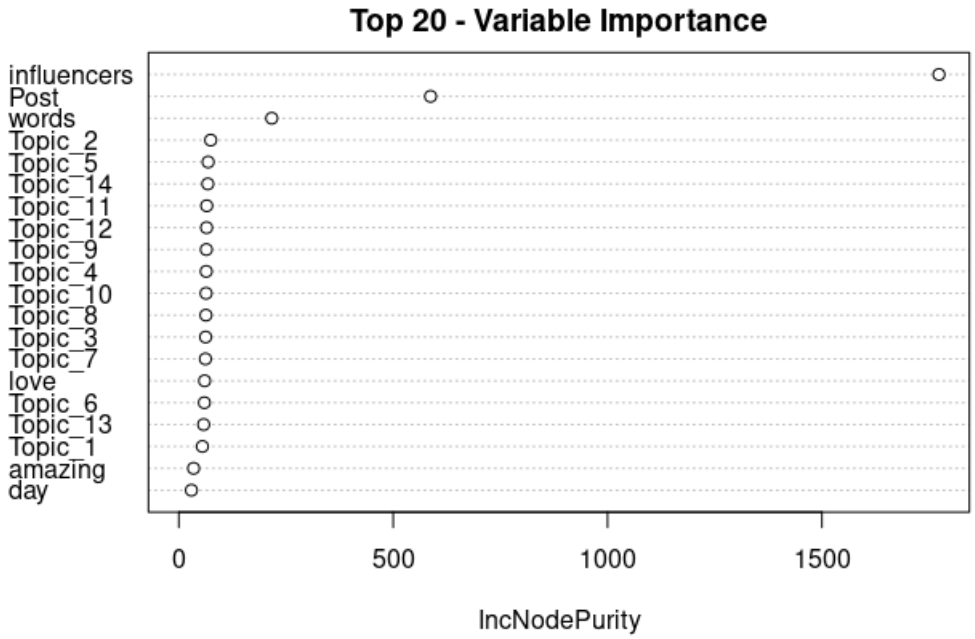
The Random Forest table (Figure 5.2) also confirms that influencers are a key variable in this study. Table 11, shows the performance of the models and random forest typically as with tree-based methods, performs better than the others, hence the evaluation was based on the Out of Bag (OOB) error, which indicates that the model performed well both on the training and test set. Thus, the random forest variable importance plot enriches this study.

Table 10: Lasso Model

Variable	Value
(Intercept)	0.0505485
Number of words	0.0001331
Emoji “yes”	0.0164911
Handle “yes”	0.0069236
Macro Influencers	-0.0814416
Nano Influencers	0.4390345
Number of posts	-0.0000139
Trust emotion score	-0.0040225
love	0.0057471
amazing	0.0085068
wow	0.0080726
cute	0.0030725
omg	0.0201148
pretty	0.0053239
delicious	0.0039525
nice	0.0100016
yum	0.0081975
perfect	0.0115102
awesome	0.0086653
girl	0.0062591
stunning	0.0067848
photo	0.0003299
shot	0.0131290
cool	0.0093452
pic	0.0026281
food	0.0015567
feel	-0.0013225
babe	0.0172455
enjoy	-0.0046022
birthday	0.0017848
gorgeous.love	0.0164848
pretty.love	0.0769699
amazing.wow	0.0650325
nice.pic	-0.0013870
sounds.delicious	0.0036192
ice.cream	0.0064351
yum.love	0.0714484
omg.yum	0.0202479
nice.shot	0.0708343
gorgeous.shot	-0.0026125
girl.love	0.1552195
Topic 1	-0.0015938
Topic 3	-0.0027383
Topic 9	0.0014597
Number of words x joy emotion score	0.0011672
Number of words x surprise emotion score	0.0011029
Nano Influencers x polarity sentiment score	0.0458894
Nano Influencers x anger emotion score	0.9561897
Nano Influencers x disgust emotion score	-0.0951497
Nano Influencers x fear emotion score	0.3309260
Nano Influencers x joy emotion score	0.4162905
Micro Influencers x sadness emotion score	-0.1102394
Number of posts x polarity sentiment score	-0.0000014
Number of posts x anticipation emotion score	-0.0000010
Number of posts x joy emotion score	-0.0000289

5.6 Random Forest

Figure 5.2: Random Forest Variable Importance



5.7 Evaluation of Prediction Models

Table 11: Prediction Model Evaluation

Model	**RMSE in-sample**	**RMSE test**
Linear Model	0.2490187	0.2310032
Lasso Model	0.2552725	0.2357004
Random Forest OOB	0.1327611	0.1977684

6 CONCLUSION

In conclusion, the aim of this study was to comparatively investigate consumer engagement on Instagram by the different influencer categories. The findings of this research are that, Nano influencers have a higher engagement rate compared to Micro and Macro. This was determined by t-tests, boxplots and the influencer categories as categorical variables in the linear models. The predictive models applied in this study consistently confirmed the results. This finding is inline with other studies that have investigated influencer marketing. The findings of this study have managerial relevance. because of their perceived authenticity

6.1 Recommendations

Firstly, as argued in the introduction, this study recommends to managers to leverage the perceived authenticity of Nano influencers for marketing campaigns particularly, in a niche markets. Given that different influencers have different skillsets, brands can determine which influencers fits their brand.

Secondly, Micro and Macro influencers have comparative advantages i.e. larger audiences, different skillsets, etc, this study recommends utilising other influencers, by emphasizing the brand authenticity be incorporated in the influencers marketing. Brand authenticity can be emphasized through brand advertising that highlights the history and virtues of the brand.

6.2 Research limitations and suggestions for further research

This research has limitations. Firstly, this study's scope could not link engagement to ROI for brands. However, in the findings of Kumar et al.2013, engagement increased brand awareness, sales and ROI. More research on ROI can enable brands better align their marketing with the ROI

Secondly, influencers were only considered for the period of time the data was gathered, however their performance could change based on other factors, thus further research on their campaigns can be informative.

Thirdly, engagement rate was determined based on the industry practice of combining likes and comments, however a separation of the two, and analysis of each separately could give better insights

Appendix A

Table 12: ANOVA test for 5 vs 20 LDA topics

5 Topics vc 20 Topics

res.df	rss	df	sumsq	statistic	p.value
57896	972.3373				
57881	971.0505	15	1.286763	5.113301	0

Table 13: ANOVA test for 20 vs 15 LDA topics

15 Topics vs 10 Topics

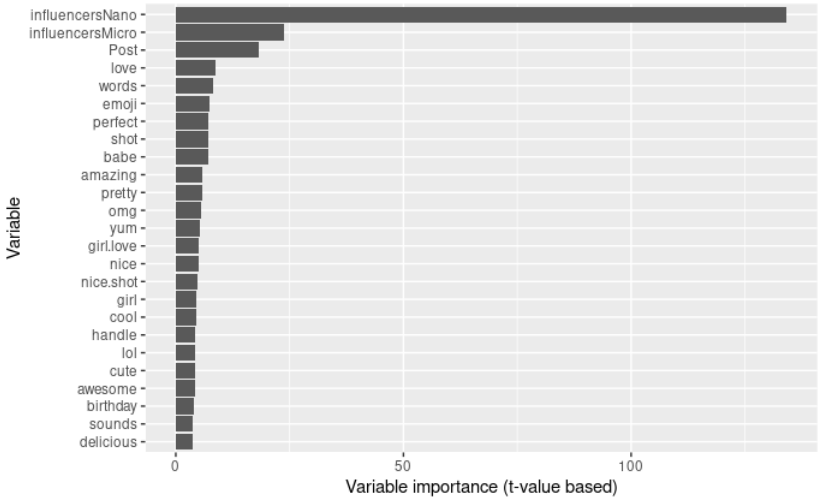
res.df	rss	df	sumsq	statistic	p.value
57891	971.7468				
57886	971.1426	5	0.6041944	7.202732	9e-07

Table 14: ANOVA test for 15 vs 10 LDA topics

20 Topics vc 15 Topics

res.df	rss	df	sumsq	statistic	p.value
57886	971.1426				
57881	971.0505	15	0.0920822	1.097741	0.359199

Figure 6.1: Linear Regression t-values Variable Importance



Appendix B

Table 15: t-test Micro vs Nano Enagement rate

Micro v/s Nano Influencers Engagement Rate	
Test statistic	99.36756
Mean Micro Influencers	0.1143928
Mean Nano Influencers	0.613522
Mean Difference	0.4991292
DF	11789.44
p value	0.000
Alternative hypothesis	two.sided
Welch Two Sample t-test	

Table 16: t-test Macro vs Nano Engagement rate

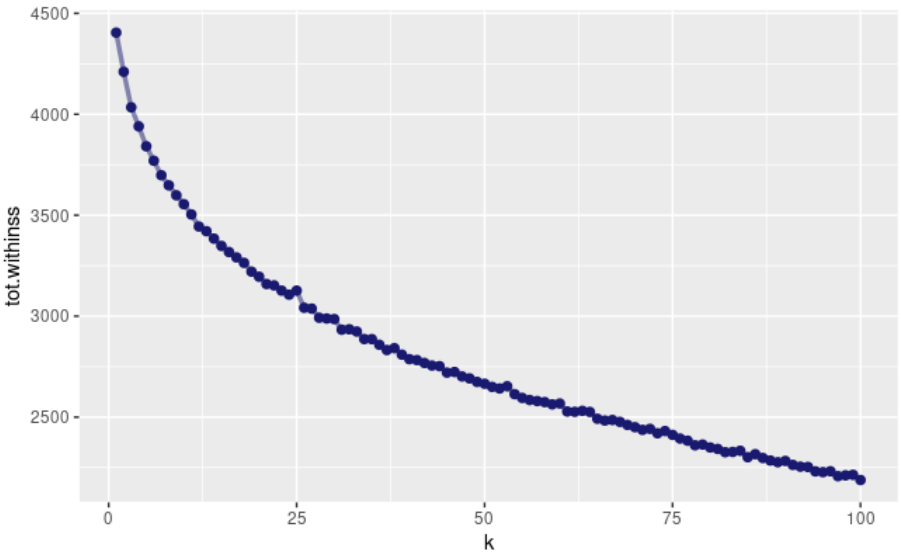
Macro v/s Nano Influencers Engagement Rate	
Test statistic	119.3963
Mean Macro Influencers	0.01632665
Mean Nano Influencers	0.613522
Mean Difference	0.5971953
DF	11591.32
p value	0.000
Alternative hypothesis	two.sided
Welch Two Sample t-test	

Table 17: t-test Micro vs Macro Enagement rate

Micro v/s Macro Influencers Engagement Rate	
Test statistic	201.1548
Mean Micro Influencers	0.1143928
Mean Macro Influencers	0.01632665
Mean Difference	0.09806616
DF	40355.63
p value	0.000
Alternative hypothesis	two.sided
Welch Two Sample t-test	

Appendix C

Figure 6.2: K-Means scree plot



Appendix D

Figure 6.3: Lasso lambda

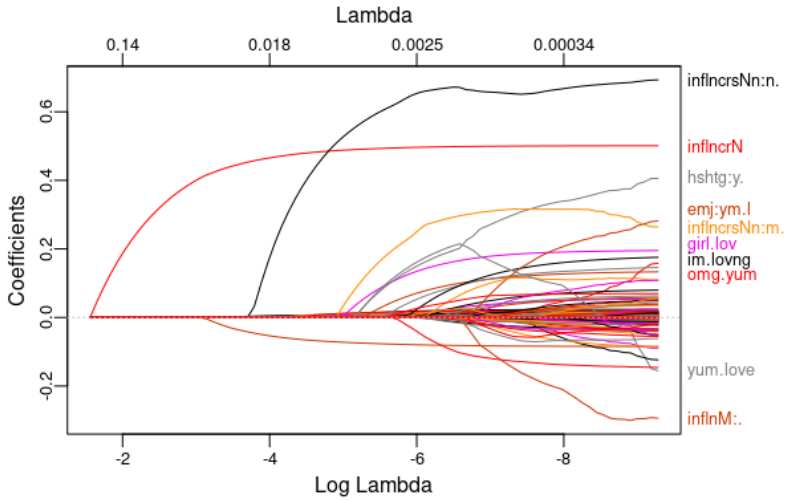
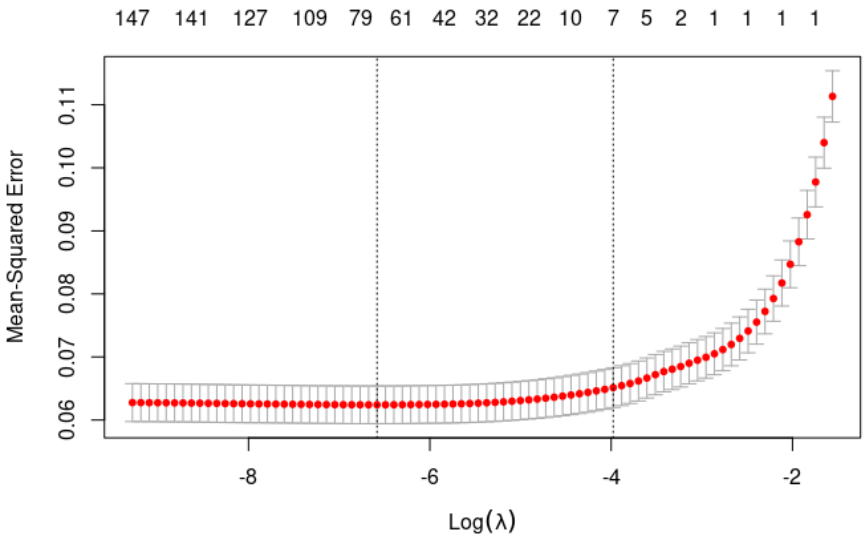


Figure 6.4: Lasso shrinkage



REFERENCES

[1] Ahmad, A. and Amin, M. R. (2016). Bengali word embeddings and it's application in solving document classification problem. In 2016 19th International Conference on Computer and Information Technology (ICCIT),pages 425–430. IEEE.

[2] Alba, Joseph W. and J.Wesley Hutchinson (1987), "Dimensions of Consumer Expertise," Journal of Consumer Research, 13 (4), 411–54.

[3] Aldous, K K An J and Jansen, B. J. (2019). View, like, comment, post: Analyzing user engagement by topic at 4 levels across 5 social media platforms for 53 news organizations. In ICWSM.

[4] ANA (2020), `https://www.ana.net/miccontent/show/id/rr-2020-state-of-influence` (accessed 16 Feb 2022).

[5] Bakalash, T. and Riemer, H. (2013), "Exploring ad-elicited emotional arousal and memory for the ad using fMRI", Journal of Advertising, Vol. 42 No. 4, pp. 275-291.

[6] Bakker, I., van der Voordt, T., Vink, P. and de Boon, J. (2014), "Pleasure, arousal, dominance: Mehrabian and Russell revisited", Current Psychology, Vol. 33 No. 3, pp. 405-421.

[7] Bakhshi S, Shamma DA and Gilbert .E.(2014) Faces engage us: Photos with faces attract more likes and comments on Instagram. Proceedings of the 32nd Annual ACM conference on Human Factors in Computing Systems. Toronto: ACM, 965–974.

[8] Bearden, William O. and Michael Etzel (1982), "Reference Group Influence on Product and Brand Decisions," Journal of Consumer Research, 9 (4), 183–194.

[9] Belanche, D., Flavián, C. and Pérez-Rueda, A. (2017), "Understanding interactive online advertising: congruence and product involvement in highly and lowly arousing, skippable video ads", Journal of Interactive Marketing, Vol. 37, pp. 75-88.

[10] Berger, Jonah and Eric Schwartz (2011), "What Drives Immediate and Ongoing Word of Mouth?" Journal of Marketing Research, 48 (5),869–80.

[11] Berger, Jonah and Katherine L. Milkman (2012), "What Makes Online Content Viral?" Journal of Marketing Research, 49 (2), 192–205

[12] Blei, D. M.; Ng, A. Y.; and Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning research 3:993–1022.

[13] Brodie, R. J., Hollebeek, L. D., Jurić, B., Ilić, A. (2011). Customer engagement: Conceptual domain, fundamental propositions, and implications for research. Journal of Service Research, 14(3), 252–271.

[14] Campbell, Colin, and Justine Rapp Farrell. 2020. "More than Meets the Eye: The Functional Components Underlying Influencer Marketing." Business Horizons 63(4) 469–79.

[15] Cho, C.-H., and Cheon, H. J. (2003). Why do people avoid advertising on the internet? Journal of Advertising, 33(3), p. 89-97.

[16] Coulter, Keith (1998), "The Effects of Affective Responses to Media Context on Advertising Evaluations," Journal of Advertising, 27(4),41.

[17] De Veirman, M., Cauberghe, V. ve Hudders, L. (2017). Marketing through Instagram influencers: the impact of number of followers and product divergence on brand attitude, International Journal of Advertising, 36(5), 798-828.

[18] Djafarova, E. and Rushworth, C. (2017), "Exploring the credibility of online celebrities' Instagram profiles in influencing the purchase decisions of young female users", Computers in Human Behavior, Vol. 68, pp. 1-7.

[19] Dokyun Lee, Kartik Hosanagar, Harikesh S. Nair (2018) Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook. Management Science 64(11):5105-5131.

[20] Evans, N.J., J. Phua, J. Lim, and H. Jun. (2017). Disclosing Instagram influencer advertising: The effects of disclosure language on advertising recognition, attitudes, and behavioral intent. Journal of Interactive Advertising 17, no. 2: 138–12

[21] Fleck, Nathalie D., Michael Korchia, and Isabelle Le Roy (2012), "Celebrities in Advertising: Looking for Congruence or Likability?" Psychology and Marketing, 29,9,651–62.

[22] Forbes 2016, `https://www.forbes.com/sites/kurtbadenhausen/2017/02/16/cristiano-ronaldo-generated-500-million-in-value-for-nike-in-2016/?sh=a4c2b88c3e94` (accessed 16 Feb 2022).

[23] Friedman, J., Hastie, T., Tibshirani, R. (2017). The elements of statistical learning (Vol. 2, No. 10). New York: Springer series in statistics

[24] Garbarino, Ellen and Mark S. Johnson. (1999). "The Different Roles of Satisfaction, Trust, and Commitment in Customer Relationships," Journal of Marketing, 63 (April): 70–87

[25] Gupta, S and Lehmann, D. R. (2005). Managing customers as investments.Upper Saddle River: Wharton School Publishing.

[26] Haenlein, M., Anadol, E., Farnsworth, T., Hugo, H., Hunichen, J., & Welte, D. (2020). Navigating the new era of influencer marketing: How to be successful on Instagram, TikTok, co. California Management Review, 63(1), 5–25

[27] Harmeling, C.M.,Moffett, J.W., Arnold,M. J., Carlson, B. D. (2017). Toward a theory of customer engagement marketing. Journal of the Academy of Marketing Science, 45(3), 312–335.

[28] Hollebeek L. (2011) , Exploring customer brand engagement: Definition and themes. Journal of Strategic Marketing, 19(7), 555–573.

[29] Hollebeek, L.D., Srivastava, R.K. and Chen, T. (2019), "SD logic–informed customer engagement: integrative framework, revised fundamental propositions, and application to CRM", Journal of the Academy of Marketing Science, Vol. 47 No. 1, pp. 161-185.

[30] Holmqvist, J. and Lunardo, R. (2015), "The impact of an exciting store environment on consumer pleasure and shopping intentions", International Journal of Research in Marketing, Vol. 32 No. 1, pp. 117-119

[31] Hudders, L., S. De Jans, and M. De Veirman. 2020. The commercialization of social media stars: A literature review and conceptual framework on the strategic use of social media influencers. International Journal of Advertising 40, no. 3: 327–375

[32] Hughes, Christian, Vanitha Swaminathan, and Gillian Brooks (2019), "Driving Brand Engagement Through Online Social Influencers: An Empirical Investigation of Sponsored Blogging Campaigns," Journal of Marketing, 83 (5), 78–96.

[33] llicic, Jasmina, and Cynthia M. Webster. 2016. "Being True to Oneself: Investigating Celebrity Brand Authenticity."Psychology Marketing 33 (6) 410–20.

[34] ITU 2021, `https://www.itu.int/itu-d/reports/statistics/facts-figures-2021/` (accessed 16 Feb 2022).

[35] Jaakonmäki R, Müller O, Vom Brocke J (2017). The impact of content, context, and creator on user engagement in social media marketing. In: Proceedings of the 50th Hawaii International Conference on System Sciences.

[36] Jamali, S. and H. Rangwala, "Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis", in: Proceedings of the International Conference on Web Information Systems and Mining, IEEE, Shanghai, China, 2009, pp. 32-38.

[37] Jin, Seung-A. Annie, and Joe Phua. 2014. "Following Celebrities' Tweets about Brands: The Impact of Twitter-Based Electronic Word-of-Mouth on Consumers' Source Credibility Perception, Buying Intention, and Social Identification with Celebrities." Journal of Advertising 43 (2):181–95.

[38] Joshi, Ashwin W. and Sanjay Sharma (2004), "Customer Knowledge Development: Antecedents and Impact on New Product Performance," Journal of Marketing, 68 (October), 47-59.

[39] Kay, Samantha, Rory Mulcahy, and Joy Parkinson. 2020. "When Less Is More: The Impact of Macro and Micro Social Media Influencers' Disclosure." Journal of Marketing Management 36 (3-4):248–78.

[40] Kelting, K., and Rice, D. H. (2013). Should we hire David Beckham to endorse our brand? Contextual interference and consumer memory for brands in a celebrity's endorsement portfolio. Psychology Marketing, 30, 602–613

[41] Kowalczyk, C.M. and Pounders, K.R. (2016), "Transforming celebrities through social media: the role of authenticity and emotional attachment", Journal of Product Brand Management, Vol. 25 No. 4, pp. 345-356.

[42] Kumar, V., Aksoy, L., Donkers, B., Venkatesan, R., Wiesel, T. and Tillmanns, S. (2010), "Undervalued or overvalued customers: capturing total customer engagement value", Journal of Service Research, Vol. 13 No. 3, pp. 297-310.

[43] Kumar V.,Vikram Bhaskaran, Rohan Mirchandani, and Milap Shah(2013), "Creating a Measurable Social Media Marketing Strategy:Increasing the Value and ROI of Intangibles and Tangibles for Hokey Pokey," Marketing Science, 32 (2), 194–212

[44] Lee D, Hosanagar K, Nair HS (2018) Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook. Management Science 64(11):5105–5131.

[45] Lee, E., Lee, J.A., Moon, J.H. and Sung, Y. (2015), "Pictures speak louder than words: motivations for using Instagram", Cyberpsychology, Behavior, and Social Networking, Vol. 18 No. 9, pp. 552-556.

[46] Lemon, K.N. and Verhoef, P.C. (2016), "Understanding customer experience throughout the customer journey", Journal of Marketing, Vol. 80 No. 6, pp. 69-96.

[47] Li and Xie (2020), Is a picture worth a thousand words? An empirical study of image content and social media engagement. Journal of Marketing Research, 57(1):1–19.

[48] Lipsman, A., Mudd, G., Rich, M. and Bruich, S. (2012), "The power of 'like'. How brands reach (and influence) fans through social-media marketing", Journal of Advertising Research, Vol. 52 No. 1, pp. 40-52.

[49] Lou, Chen and Shupei Yuan (2019), "Influencer Marketing: How Message Value and Credibility Affect Consumer Trust of Branded Content on Social Media," Journal of Interactive Advertising, 19(1), 58–73.

[50] Machado, P., Romero, J., Nadal, M., Santos, A., Correia, J. and Carballal, A. (2015), "Computerized measures of visual complexity", Acta Psychologica, Vol. 160, pp. 43-57.

[51] McCracken, Grant (1989), "Who Is the Celebrity Endorser? Cultural Foundations of the Endorsement Process," Journal of Consumer Research, 16 (3), 310–21.

[52] Mehrabian, A. and Russell, J.A. (1974), An Approach to Environmental Psychology, MIT Press, Cambridge, MA.

[53] Moorman, C., Deshpande, R., Zaltman, G. (1993). Factors affecting trust in market research relationships. Journal of Marketing, 57(1),81–101.

[54] Morhart, Felicitas, Lucia Malär, Amélie Guèvremont, Florent Girardin, and Bianca Grohmann. 2015. "Brand Authenticity: An Integrative Framework and Measurement Scale." Journal of Consumer Psychology 25(2), 200–18.

[55] (MSI Practice Winner 2011-2012), Hokey Pokey – Gary L. Lilien ISMS-MSI Practice Prize Videos (lilienpracticeprizevideos.org) (accessed 26 February 2022).

[56] Mudambi, Susan M. and David Schuff (2010), "What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com," MIS Quarterly, 34 (1), 185-200.

[57] Pansari, Anita and V. Kumar (2017), "Customer Engagement: The Construct, Antecedents, and Consequences," Journal of the Academy of Marketing Science, 45 (3), 294–311.

[58] Park, Jiwoon, Ji Min Lee, Vikki Yiqi Xiong, Felix Septianto, and Yuri Seo. 2021. "David and Goliath: When and Why Micro-Influencers Are More Persuasive than MegaInfluencers." Journal of Advertising 50 (5):584–602.

[59] Pennington J., R. Socher, and C. Manning. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543, 2014.

[60] Pereira, H. G., Salgueiro, M. F., and Mateus, I. (2014) Say yes to Facebook and get your customers involved! Relationships in a world of social networks. Business Horizons, 57, 695-702.

[61] Petty, Richard and John Cacioppo (1986), "The Elaboration Likelihood Model of Persuasion," Advances in Experimental Social Psychology, 19 (12),123-205.

[62] Pew Research Center 2021, `https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/` (accessed 28 April 2022)

[63] Prahalad, Coimbatore K. and Venkat Ramaswamy (2004), "Co-Creation Experiences: The Next Practice in Value Creation,"Journal of Interactive Marketing, Vol. 18, 5-14.

[64] Racherla, Pradeep, Munir Mandviwalla, and Daniel J. Connolly (2012),"Factors Affecting Consumers' Trust in Online Product Reviews," Journal of Consumer Behaviour, 11, 2, 94–104.

[65] Richins, Marsha L. (1997), "Measuring Emotions in the Consumption Experience," Journal of Consumer Research, 24 (September), 127–46.

[66] Risch J and R. Krestel, "Top comment or flop comment? predicting and explaining user engagement in online news discussions," in Proc. Int.Conf. on Web and Social Media, 2020, pp. 579–589.

[67] Rooderkerk, Robert P. and Koen H. Pauwels (2016), "No Comment?! The Drivers of Reactions to Online Posts in Professional Groups," Journal of Interactive Marketing, 35, 1–15.

[68] Ryu, Gangseog and Lawrence Feick (2007), "A Penny for Your Thoughts: Referral Reward Programs and Referral Likelihood," Journal of Marketing, 71 (1), 84.

[69] Santini, F., Ladeira, W. J., Pinto, D. C., Herter, M. M., Sampaio, C. H., Babin, B. J.(2020). Customer engagement in social media: A framework and meta-analysis. Journal of the Academy of Marketing Science, 48(6), 1–18.

[70] Shopify 2021, https://www.shopify.com/blog/instagram-influencer-marketing/ (accessed 26 April 2022)

[71] Socialpilot 2021, https://www.socialpilot.co/instagram-marketing/instagram-stats/ (accessed 28 April 2022)

[72] Statista 1 2021, https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/ (accessed 16 Feb 2022).

[73] Statista 2 2021, https://www.statista.com/statistics/421169/most-followers-instagram/ (accessed 10 Feb 2022).

[74] Suh, B., L. Hong, P. Pirolli, and E. H. Chi, "Want to Be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network", in: Proceedings of the Second International Conference on Social Computing, IEEE, Minneapolis, MN, 2010, pp. 177-184.

[75] Taylor, C.R. (2009), "The six principles of digital advertising", International Journal of Advertising, Vol. 28 No. 3, pp. 411-18.

[76] Tirunillai and Tellis (2014),"Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation," Journal of Marketing Research, 51 (8), 463–79.

[77] Trusov, Michael, Randolph E. Bucklin, and Koen Pauwels (2009), "Effects of Word-of-Mouth Versus Traditional Marketing: Findings from an Internet Social Networking Site," Journal of Marketing, 73, 5, 90–102

[78] Uribe, Rodrigo, Cristian Buzeta, and Milendka Vela´squez (2016), "Sidedness, Commercial Intent, and Expertise in Blog Advertising," Journal of Business Research, 69 (10), 4403–10.

[79] US census 2020, https://www.census.gov/quickfacts/fact/table/US/PST045221 (accessed 28 Feb 2022).

[80] US commission 2017, https://www.ftc.gov/business-guidance/resources/ftcs-endorsement-guides-what-people-are-asking/ (accessed 27 April 2022)

[81] Van Doorn, J., Lemon, K. N., Mittal, V., Nass, S., Pick, D., Pirner, P., Verhoef, P. C. (2010). Customer engagement behavior: Theoretical foundations and research directions. Journal of Service Research, 13(3), 253–266.

[82] Van Reijmersdal, E.A.; Neijens, P.C. and Smit, E.G. (2010). How Media Factors Affect Audience Responses to Brand Placement. International Journal of Advertising, 29(2), 279-302.

[83] Verhoef, Peter C., Werner Reinartz and Manfred Krafft, (2010), "Customer Engagement as a New Perspective in Customer Management," Journal of Service Research, 13 (3), 247-252.

[84] Vieira, V.A. (2013), "Stimuli-organism-response framework: a meta-analytic review in the store environment", Journal of Business Research, Vol. 66 No. 9, pp. 1420-1426.

[85] Vivek, S. D., Beatty, S. E., Morgan, R. M. (2012). Customer engagement: exploring customer relationships beyond purchase. Journal of Marketing Theory and Practice, 20(2), 122–146.