UiT The Arctic University of Norway

Fakultet for biovitenskap, fiskeri og økonomi

Exploring quantitative modelling of semantic factors for content marketing

Kevin Xavier

Masteroppgave i Økonomi og Administrasjon BED-3901-1 Vår 2023

Foreword

This master's thesis is the result of two years of study at UiT in the field of Business and Administration.

First and foremost, I would like to thank my main advisor Nils Magne Larsen for his guidance, excellent communication, and honest and direct feedback that has shaped this thesis into what it is today.

I would also like to thank the UiT faculty at Handelshøgskolen. They have laid the foundation for the thesis both in terms of developing me throughout my studies but also by sowing the early seeds that led me down the path of the topics discussed in this thesis.

Lastly, I would like to thank my fiancée for her relentless support, pointed criticism, and for bringing me to Tromsø in the first place. Without her I would be a different person and this thesis would not exist.

Abstract

Developments in business analytics as well as an increased availability of data has allowed digital marketers to better understand and capitalize on consumer behavior to maximize the engagement with marketing materials. However, because most previous studies in this field have focused on consumer behavior theory, they have been largely limited in scope due to small datasets and reliance on human-labeled data. This study aims to explore the potential of using a machine-learning language model to generate vector embeddings, representing the semantics in text, to model engagement in a quantitative way. By clustering the semantic vector embeddings, the study was able to generate datasets on different topics, on which regression models were estimated to gauge the impact of the represented variables. Many of the parameters in the models were shown to be significant, implying both explanatory potential in text semantics, as well as the presented methods' ability to model these. This expands on theories in the literature regarding how semantic factors affect consumer perception, as well as highlighting that text semantics contains information that can help inform marketing decision-making. The paper contributes a methodology that can allow academics and marketers alike to model these semantics and thus gain insights relating to how topics and language affect consumer engagement. Further investigation into similar methods might allow digital marketers to improve their understanding how different consumers perceive and engage with their marketing content.

Keywords: "social media marketing", "topic modelling", "text embedding", "user generated content", "marketing analytics"

Table of Contents

List of Figures

List of Tables

1 Introduction

Recent developments in business analytics have had a significant impact on both the academic study and implementation of business operations and strategies (Saura, 2021). This includes increased availability of data, widespread adoption of large-scale computational infrastructure, and the development of new machine learning techniques such as advanced Natural Language Processing (NLP) and high-dimensional clustering (Wedel & Kannan, 2016). Digital marketing is a field that has been particularly affected by this revolution, as it has both been blessed with exceptional availability of data and cursed with a multitude of abstract business problems. This is particularly prominent with respect to the development of ads and interaction with customers through social media (Lies, 2019).

Social media is a general term for a group of internet applications that allow users to share content in a variety of different ways (Chi, 2011). These platforms allow businesses to communicate effectively with their potential customers, as well as for customers to communicate between each other through User Generated Content (UGC) (Kaplan & Haenlein, 2010). Because of its nature, social media has become a focus of many companies' marketing strategies, but research has shown that many still struggle to utilize it effectively or lack a strategic foundation for it (de Oliveira Santini, et al., 2020; Paruthi & Kaur, 2017; Holmes, 2015; de Clerck, 2013). A better understanding of how consumers interact with and perceive content on social media therefore has huge potential value in terms of both producing better marketing but also developing more effective marketing strategies.

Social Media Marketing has proven beneficial to businesses in a variety of different ways. It provides new touchpoints for businesses to reach shoppers (Venkatesh, et al., 2011) and through this interaction improves customer perception and brand awareness, leading to a better brand image, more trust, and loyalty (Alalwan, et al., 2017; Bianchi & Andrews, 2018; Zhang, et al., 2018; Godey, et al., 2016; Brodie, et al., 2013; Gummerus, et al., 2012; Zheng, et al., 2015). Studies have found social media to increase customer involvement and improve customer relationship management (Alalwan, et al., 2017), as well as create purchase intentions (Alalwan, et al., 2017; Kumar, et al., 2013; Cheung, et al., 2012; Kumar, et al., 2010; Beukeboom, et al., 2015) and positive word of mouth, leading to customer evangelization (Alalwan, et al., 2017; Kumar, et al., 2013; Cvijikj & Michahelles, 2013).

Social media has been found to be a more cost-effective marketing channel than traditional media (Leeflang, et al., 2014), and utilizing it can therefore lead to improved financial performance (Kim & Chae, 2018; Cvijikj & Michahelles, 2013), and accelerate a firm's value (Kim, et al., 2015). Use of marketing on social media can also correlate with improved effectiveness of other forms of advertising (Lin, et al., 2018) as well as increased sales value (Kumar, et al., 2017). Many studies also emphasize how social media has changed the way customers gather information, with it increasingly becoming a main source for many (Alalwan, et al., 2017; Hamilton, et al., 2016). Information from social media is also seen as more reliable and less biased than information coming from a company directly, which might help make it even more influential as a marketing tool (Alves, et al., 2016).

On social media, users engage with the available content based on what they find interesting or feel like they derive value from (Brodie, et al., 2011). Consumer engagement in social media posts is expressed in the form of liking, commenting, or sharing posts, as well as reaching out to other people thought the posts (Pino, et al., 2019). Intensive engagement has been clearly correlated with increased revenue as well as improved underlying brand perception (Cvijikj & Michahelles, 2013; Gutiérrez-Cillán, et al., 2017). The marketing literature has therefore identified engagement to be the most important metric when it comes to social media marketing success, and content to be the most important determinant of it (Peruta & Shields, 2018; Brodie, et al., 2013; Santos, et al., 2021; Lin, et al., 2018; Parent, et al., 2011; Vrontis, et al., 2021; Farook & Abeysekara, 2016). However, most previous studies have approached modelling of engagement through a consumer behavior lens (Santos, et al., 2021; Paquette, 2013; Alves, et al., 2016). Previous studies have also been very limited in scope, focusing on small niches, using hand-labelled datasets, and often producing contradictory findings (Gutiérrez-Cillán, et al., 2017; Schultz, 2017). At the same time, multiple calls for future research have emphasized the need for a better understanding of creating engagement and practical application of theoretical principles (Kujur & Singh, 2018; Alves, et al., 2016; Aydinn, 2019).

The recent advancements in business analytics might be exactly what is needed to provide this expanded understanding. With the help of machine-learning based language models, text data such as words or sentences can be mapped to meaningful quantitative representations such as text vector embeddings (Fang, et al., 2022; Mikolov, et al., 2013b). These methods have seen applications in social science research but have not yet made it do the field of

content marketing where they could allow researchers to extract and model new types of variables from social media data, such as topics or meaning within text, on a large scale (Peruta & Shields, 2018; Pino, et al., 2019; Viswanathan, et al., 2018). As these embeddings represent the semantics within texts, they might hold significant explanatory potential with regards to online consumer engagement.

This study aims to conduct an exploratory study into a novel method of utilizing text vector embeddings extracted with a language model to both segment social media content into different topics, as well as measure the effect of vector represented semantic factors on consumer engagement.

The study's goals are to:

- *Assess the potential of using document vector embeddings to represent content semantics and model consumer engagement.*
- *Gauge how the impact of the vector embeddings and other semantic-related variables changes depending on the topic of the content.*

The main contribution of this study is methodological, as it aims to provide a framework for a novel approach to model engagement with respect to factors that the academic literature has not previously been possible to take into account. The outcome of this study could have implications for academics seeking to understand consumer engagement by providing them with a method that can gather significant amounts of data for specific topics, as well as account for variables that have previously not been studied. In this way the study might help remove some of the limitations that has plagued the field, such as reliance on hand-labelled data and small datasets, and could provide a new perspective on the contradictory findings that might result from these. It would also allow future studies to use the method to develop a deeper understanding of the impact of semantics on engagement, which would lead to a better understanding of consumer behavior and values. Down the road, this kind of research could allow businesses to both better understand and execute on social marketing strategies and thus derive more value from their operations.

However, the scope of the study is limited regarding some aspects. It does not intend to explore differences in engagement between different social media platforms, and its findings can therefore not necessarily be applied to these. This also a limitation with regard to other types of marketing in other types of media. The study is also severely limited in terms of

what type of content is analyzed, since it does not consider the picture or video that is often the main component of a post on social media, but instead relies solely on the text describing it. Lastly, the study only looks at social media posts from a specific time period, and although time is controlled for, the findings do not necessarily apply to other time periods as well.

Chapter two of this study aims to provide an overview of existing literature and theories relevant to it. It first covers social media marketing in general, and then dives further into conceptual and empirical studies. Literature and concepts relevant to the paper's specific method are then covered, and a conceptual model is proposed that illustrates the relationships between variables that the study aims to explore. The paper's methodological framework is then laid out in chapter three, after which results from the data and analysis are presented in chapter four. In chapter five the results are discussed and compared against the goals of the study, which includes potential implications of the findings as well as suggestions for future research and improvements to the methodology. A conclusion is written in chapter six that summarizes the study's contribution.

2 Literature Review

A literature review was conducted in order to establish a theoretical foundation for the research, as well as provide context around its relevance and possible consequences. Information for the review was gathered by first identifying and reviewing existing literature relating to social media marketing, content marketing, and factors affecting them. Five relevant systematic literature reviews were found and used as basis: (Paquette, 2013; Alves, et al., 2016; Alalwan, et al., 2017; Santos, et al., 2021; Vrontis, et al., 2021) .

A search into existing literature was then done following the systematic method in Hamari, et al., (2014), by first searching the Google Scholar database using the terms “social media content”, “social media engagement”, and “social media factors”. Results were then filtered to only include articles that were peer reviewed and directly relevant to the topic studied. Although conducted, backwards reference research generally had little success because of the novelty of the subjects.

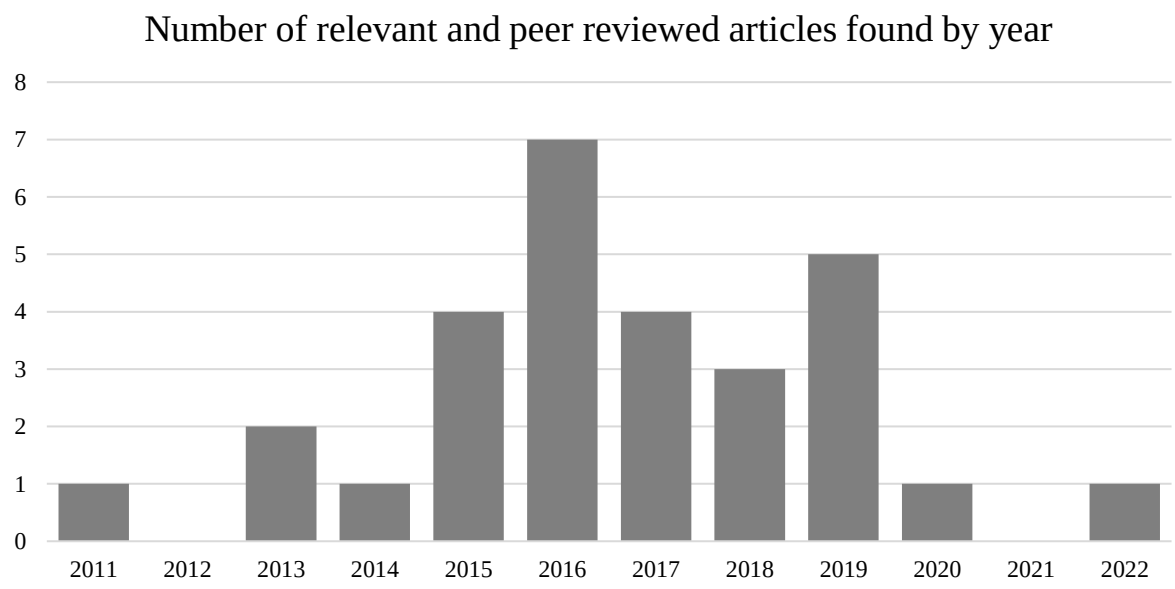

Figure 1. Number of relevant and peer reviewed articles by year

The method yielded 29 articles that were read and analyzed, and the relevant information was synthesized into the review below. The review is structured to first provide a broad overview of social media marketing as a concept, and then go further in depth into content marketing specifically. Under content marketing, relevant conceptual studies and theories are covered, as well as empirical studies and their theoretical foundation. Language modelling is then

covered in order to provide a foundation for the methodology, and lastly a conceptual model is presented.

2.1 Social media marketing

Social media as an academic concept in the literature is relatively new and it does therefore not have an established definition (Guesalaga, 2015). Some, such as Marketo (2010) define it based on a literal description as the exchange of information and interactions on online platforms. It has also been defined in terms of its functionality, such as by Kietzmann et al., (2011) that states that social media should be described based on its user's identity, presence, sharing, relationships, groups, conversations, and reputations. Social Media Marketing literature tends to refer to social media as the specific platforms on which the communication between users happens, and where businesses therefore can exert influence on potential customers (Guesalaga, 2015).

As a marketing channel, social media is useful based on its ability to reach the right audiences with the right messages (Dwivedi, et al., 2015). Large amounts of user data, as well as machine learning algorithms allow marketing on the platforms to be focused on specific target audiences to an extent not seen before, as well as find potential new segments to target based on previous campaigns (Santos, et al., 2021). Messages and information can also be easily changed and adapted to different audiences instantly based on A/B tests and real-time conversion rates.

2.1.1.1 Content Marketing

Marketing on social media allows for multiple different types of strategies ranging from organic inbound marketing (building audiences through engaging content) to paid advertising (using a platform directly as an advertising channel) (Lee, et al., 2015; Peruta & Shields, 2018). Of these two, organic inbound marketing has received the most focus in research as it differs the most from conventional marketing and is the least understood. Developing brand related content, called Content Marketing (CM), is key to this kind of organic marketing and relies on engaging customers in order to generate positive brand associations (Lei, et al., 2017). Content, with regards to CM, is defined as static or rich media posted online that aims to educate or compel audiences (Holliman & Rowley, 2014). A modern definition of CM by Asmussen et al., (2015) broadens the concept to any media produced by a brand's stakeholders with primary aim of generating digital engagement relating to a brand or its products.

Digital engagement is derived from the consumer behavior concept of Customer Engagement (CE) which was pioneered in the digital marketing literature by Brodie et al., (2011) as a driver of customer loyalty from their interaction with marketing media. It has since developed to describe the intensity by which a user connects with the content, and includes a multitude of marketing consequences such as increased brand exposure, revenues, and profits (Cvijikj & Michahelles, 2013; Gutiérrez-Cillán, et al., 2017). Engagement is therefore central within content marketing literature as a critical mediator variable between content and beneficial marketing outcomes, and it is often used as a key metric to determine the success of content marketing (Lei, et al., 2017; Kumar & Nayak, 2018; Kumar & Kumar, 2019).

The literature on content marketing generally comes in one of two forms: conceptual studies based on theories previously developed in consumer behavior literature, and empirical studies that research the cause-and-effect relationship between social media usage and marketing outcomes (Paquette, 2013; Guesalaga, 2015; Santos, et al., 2021).

2.1.2 Conceptual studies

Several academic theories have been used to provide a framework for, and help understand, the mechanisms behind social media as a marketing tool. These stem from the fields of consumer psychology and consumer behavior, and aim to explain consumers motivations and internal reasoning. Two theories of particular note to content marketing literature are the Uses and Gratification Theory (UGT) and Social Identity Theory (SIT) (Santos, et al., 2021). These are used by researchers to develop hypotheses with respect to factors that affect social media marketing outcomes as they lay a foundation for the consumers perception of the media (Leung & Tanford, 2015).

2.1.2.1 Uses and Gratification Theory

Uses and Gratification Theory was first developed by Katz et al., (1973) and aims to explain the link between Maslow's hierarchy of needs and how people use media by attributing their interactions with it to the use and gratification they expect to get from it. It was developed to help understand mass communications from the audience's perspective and was groundbreaking as the first theory to model the audience as active rather than passive (Blumler, 1979). The theory is based on the assumption that different audiences perceive and interact with different kinds of media in different ways and outlines four factors that affect

the relationship: social circumstances, personality of the consumer, their patterns of media consumption, and the qualities of the media itself (Blumler, 1979).

As the theory is a general framework for consumer interaction with media, it also provides insight into the motivations of social media users. UGT is in fact specifically relevant to social media as it is a medium that is intrinsically interactive and allows its consumers a large degree of power over their consumption of it (Kaplan & Haenlein, 2010). Because of this, many recent studies have analyzed marketing in social media through the lens of UGT, such as Menon (2022) that identified seven different motivations behind Instagram Reels usage within a set of users. Dolan et al., (2016) also cements UGT further in the context of content marketing by developing a model for stimulating engagement based on factors in the theory. In this sense the study links consumer behavior theory with the psychological theory behind CE.

A common criticism of UGT, however, is its sole focus on one-way communication, and that it therefore might not represent modern modes of mass communication that allow individuals within the audience to respond back to the firm conducting the marketing (Baldus, et al., 2015).

2.1.2.2 Social Identity Theory

Social Identity Theory was first developed by Tajfel (1978) as a way to model social group interactions and to understand the motivations behind intergroup behaviors. The theory states that people identify with different social groups, from which they both derive their own self-identity and are judged by others. Each group also comes with an attached value and emotional significance that affects how the group and its members are perceived.

SIT consists of three parts: categorization, identification, and comparison. Categorization refers to the intrinsic categorization that we humans do in order to understand the world. In a social context this entails categorizing humans into groups depending on a multitude of demographic and social factors. Identification refers to the process by which someone feels like they become a member of the group. Social identity is identified through three dimensions: cognitive, emotional, and evaluational. Cognitive social identification is the cognitive awareness of the person belonging to a group, emotional social identification is the emotional sense of involvement to the group, and evaluational refers to the positive or negative value attached to the group (Leung & Tanford, 2015). Comparison is the last step of

the SIT process, which involves comparing groups and evaluating people based on their belonging to the groups.

SIT is used in social media marketing literature because belonging to groups is a core part of social media and SIT therefore helps researchers better understand actions and perceptions in that environment (Santos, et al., 2021). This is exemplified by the study by Kumar & Nayak (2018) which modeled brand community relationships and identification with respect to engagement. Brand community engagement has also been shown to lead to positive marketing outcomes such as brand loyalty and commitment. These studies have been instrumental in developing our understanding of how social identity, engagement, and marketing outcomes are related.

2.1.3 Empirical studies

Empirical studies in the field of CM have largely aimed to explore industry impact by modelling engagement with respect to the content that generated it as well as its context. Although engagement has been empirically found to have a multitude of positive business consequences, there is still debate regarding which factors have the most impact on it (Alalwan, et al., 2017; Aydinn, 2019). For this reason, innovation and novelty in the field usually comes from either the specific data used (source, collection method, user group), or the specific independent variables analyzed. Factors considered tend to belong to one of the following groups: consumer behavior, content value, content characteristic, content environment, or content semantic.

2.1.3.1 Engagement as a metric

Engagement in these studies is usually defined as consisting of three facets – cognitive, affective, and behavioral, that are activated in the consumer based on the content (Aydinn, 2019). In practice, engagement is measured by the number of likes, dislikes, shares, comments, or reposts that content gets, with different weights given to each depending on the platform and the method used in the study (Aydinn, 2019; Peruta & Shields, 2018; Lei, et al., 2017; Wallace, et al., 2014).

2.1.3.2 Consumer behavior factors

Some of the most studied factors in content marketing are factors determined by the consumer and their context. These are the closest factors to conventional consumer behavior and therefore tend to be modeled using similar theories and frameworks, most often SIT.

Some of these are focused on the consumer itself, such as cognitive, affective, and behavioral factors (Dessart, et al., 2016; Paruthi & Kaur, 2017), the customer's needs, motives, goals (Felix, et al., 2017), or perceived intentions (Gutiérrez-Cillán, et al., 2017), while others focus on factors external to the customer such as their relationships, social factors (Kujur & Singh, 2018), or brand relationships and communities (Kumar & Kaushik, 2018). Although highly theoretically impactful, these studies have had limited practical application as the factors are generally very abstract and difficult to measure (Alves, et al., 2016).

2.1.3.3 Content value factors

Content value factors relate to the inherent value that the content itself provides the consumer. These variables have their theoretical foundation in UGT, and usually approach the topic from the perspective that providing the customer more value correlates to higher engagement and thus marketing success. Although some of these studies have focused on a single value appeal such as entertainment value (Cvijikj & Michahelles, 2013), others have explored differences between types of value such as hedonic vs utilitarian value (Hughes, et al., 2019; Gutiérrez-Cillán, et al., 2017), and informational vs emotional appeal (Cervellon & Galipienzo, 2015; Thongmak, 2015; Menon, et al., 2019). Content value factors effectively provide a framework for strategic decision making with regards to CM, but might not always be a good guide for actual content production as providing high-value content might not always lead to positive marketing outcomes (Gutiérrez-Cillán, et al., 2017).

2.1.3.4 Content characteristic factors

Many studies have also investigated the effect of content characteristic factors on engagement as these factors relate to what the consumer of the content sees and might therefore have an impact on the consumer-behavior factors previously outlined. These studies are therefore still based on the consumer behavior theories but treat them as latent or mediating variables instead of independent. The variables used to represent post characteristics range from fully descriptive and objective such as content length (Schultz, 2017; Ibrahim, et al., 2017), and type of media (Coelho, et al., 2016), to subjective and abstract such as vividness (Aydinn, 2019; Pino, et al., 2019; Schultz, 2017; Menon, et al., 2019), or interactivity (Schultz, 2017; de Vries, et al., 2012; Menon, et al., 2019). Although the objective factors are relatively firmly established, there is still significant debate about how to define the subjective factors. Vividness, for instance, has both been defined as the representational richness of content (Lei, et al., 2017), as well as whether the post uses static vs dynamic visuals (Aydinn, 2019).

2.1.3.5 Content environment factors

The environment around the content has also been included in some studies as independent variables. These have largely focused on timing (Cvijikj & Michahelles, 2013; Peruta & Shields, 2018; Schultz, 2017) ranging from time of day to day of the week. From a theoretical perspective, timing affects the engagement of content based on how many people are shown the content, and consensus seems to be that posting during times of high activity is preferred. Uniqueness is another environmental factor that has been considered and is defined as the presence of characteristics that draw attention based on their distinctiveness (Pino, et al., 2019; Huertas & Marine-Roig, 2016). Although a descriptive characteristic, uniqueness is considered an environmental factor as it is a content's environment that determines whether it is unique or not. A post might therefore be unique in one environment but not in another one (i.e., different time of year, topic, or platform).

2.1.3.6 Content-semantic factors

Content semantic factors relate to the semantics or meaning of the produced content. Being conscious of meaning or semantics in marketing is nothing new, going back as far as Schutte (1969), but the digitalization of media and improvements in natural language processing has changed the way we are able to analyze and adjust for these factors (Li & Bond, 2022). Semantic factors are important because they affect both how consumers of media perceive a message as well as the association between brands, content and external factors relating to the brand (Hatzivassiloglou & McKeown, 1997). This is based on the assumptions that consumers attach meaning to individual words as well as adjectival expressions used by authors within text (Aggarwal, et al., 2009).

Bruce & Wiebe (1999) as well as Hatzivassiloglou & Wiebe (2000) were the first to implement a methodological approach based on these theories by associating specific adjectives with particular brands and inferring relationships between the two. This was later expanded using lexical semantics in order to develop an overview of brand positioning in an online market based in search data by Aggarwal et al., (2009).

In recent years, methodological implementations of content-semantic factors have expanded in terms of scope. One study by Peruta & Shields (2018) was found that explored the engagement in social media content relating to different topics and discussed subjects, while Pino et al., (2019) is a study that included sentence style and emotionality in its model. Sentiment has also been a focal point of much research, such as by Viswanathan et al., (2018)

that explored sentiment and engagement in UGC in relation to brand actions. A more general version of semantic similarity has also been used to model the engagement that news articles receive on social media (Li & Bond, 2022).

Although there is limited previous research within marketing on these kinds of factors, the literature seems united in its opinion that they need to be studied further and might hold significant explanatory potential with regards to engagement (Lei, et al., 2017; Peruta & Shields, 2018; Thongmak, 2015).

2.2 Language Modelling

Recent developments in the field of machine learning, specifically with regards to advancements in language modelling, present a great opportunity for marketing-researchers to develop their understanding of how language influences consumers and their interaction with media.

2.2.1 Statistical Word Embedding

A core concept within the field of NLP are statistical word embeddings, as they are integral to making the qualitative and unstructured data (words and sentences), be interpretable by a quantitative machine learning model. Word embedding models are models whose sole purpose it is to represent words in meaningful ways that preserve as much semantic information as possible (Mikolov, et al., 2013a). The development of these came about as a consequence of the Distributional Hypothesis, which assumes that words that appear in similar contexts tend to share semantic similarity (Sahlgren, 2008; Devlin, et al., 2019). Modern word embedding models are therefore trained on a continuous stream of words, and many models such as Linear Discriminant Analysis (LDA) have fallen out of favor because of their inability to capture semantic information in their embeddings (Mikolov, et al., 2013b).

In a statistical word embedding model, each word is represented by a corresponding high-dimensional vector, the shape of which is attributed to the contexts in which the words appear in the training data. The preservation of semantic information in these embeddings allow for algebraic operations to be conducted with the vectors in order to extract information. For instance, if the vector(“man”) is subtracted from the vector(“king”) and added to the vector(“woman”) the result would be a vector similar to the vector(“queen”). (Mikolov, et al., 2013b)

2.2.2 Short Text Topic Modelling

In the field of machine learning, topic modelling refers to methods that are able to infer coherent latent topics within textual data (Chen, et al., 2023). Short text topic modelling (STTM) is a subfield of topic modelling that aims to perform this task on shorter documents (Qiang, et al., 2020). This is significantly more difficult than conventional topic modelling because of the lower amount information within each document that is clustered. However, because of the increased occurrence of short text in recent years, specifically from social media, STTM has received increasing amounts of attention with a multitude of new methods and approaches (Albalawi, et al., 2020).

One such approach is the Topic2Vec model by Niu & Dai (2015), which uses a conventional clustering algorithm: Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), as well as the dimension reduction algorithm: Uniform Manifold Approximation and Projection (UMAP), to cluster document vector embeddings. These document vector embeddings are developed in a similar way to the word vector embeddings previously described but embed an entire document in a single vector instead of the individual words. This can be achieved with the help of either a Doc2Vec model (Mikolov & Le, 2014) that learns a paragraph vector alongside individual word vectors, or a sentence transformer model such as BERT (Devlin, et al., 2019).

Although evaluating topic models against each other is difficult, Topic2Vec is able to achieve comparable Normalized Pointwise Mutual Information (NPMI) scores to other models while also sporting several additional benefits (Niu & Dai, 2015). For instance, as a result of using HDBSCAN the Topic2Vec model does not need to know in advance how many topics to create, which allows for both faster training and iteration as well as implementation on less understood datasets such as Big-Data from social media. As an embedding-based model, it also generates fewer “useless” topics (Harrando, et al., 2021). The topics are also embedded in the same semantic vector space as the word embeddings, which allows for representing topics with semantically similar words in an efficient and systematic way. Topic2Vec is also specifically good at modelling short texts, giving the model a distinct advantage for digital marketing use cases (Niu & Dai, 2015).

There are many examples of marketing studies that use topic modelling in order to extract and categorize information from digital data sources. By using a topic modelling algorithm

on online reviews Korfiatis et al., (2018) was able to find the dominant drivers of customer satisfaction in a specific hotel industry leading to a better understanding on niche positioning in the market. A study by Gregoriades et al., (2021) used a similar topic modelling technique in order to analyze electronic word of mouth in order to improve communication to customers. Topic modelling can also be used to understand consumers intentions, such as by Wang et al., (2022) that modelled user behavior based on semantics in social media data, and Luo et al., (2020) that used topic modelling to gain insight into theme park visitor perceptions based on reviews. However, no existing articles were found that utilize topic modelling techniques directly in relation to engagement.

2.3 Conceptual Model

Figure 2 is an outline of the relationships between the variables that this study aims to research. Independent variables were chosen based on a mix of previous research within content marketing and previous marketing research that uses semantic variables, as well as the post text as represented by a vector embedding. The objective of the model is to estimate the impact of these semantic variables on engagement, which was therefore chosen as the dependent variable. As the study also aims to investigate the differences between topics, post topic is present in the model as a moderating variable. The reasoning for this is that different topics and the communities that discuss those topics online, are interested in and value different things, which might therefore affect the importance placed on other content-related variables.

A more detailed description of the variables included in the model and how they were measured is provided in the method section.

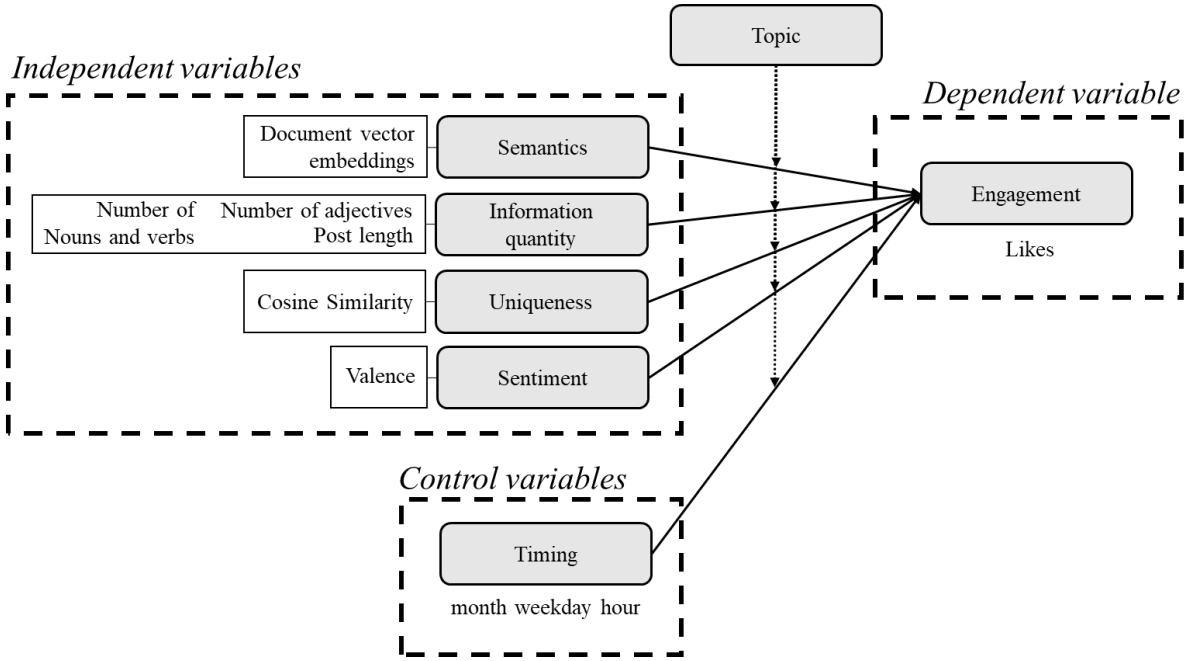

Figure 2. Proposed conceptional model for relationships between independent and dependent variables.

3 Method

In the following chapter the study's methodological framework will be presented in depth. Choice of research design will be justified and related back to the goals of the study – *to assess the potential of using document vector embeddings to model engagement, and to gauge how the impact of these and other semantic variables changes depending on topic*. Specific methodology will then be outlined, which aims to give an account of the study's source and use of data, variable extraction, and modelling. Each of these will be examined in terms of their relevance as well as their impact on the potential generalizability of the study's findings.

3.1 Overview of research design

The method was designed around two key criteria for meeting the study's goals:

- Achieving the highest possible quality text embeddings
- Producing topic clusters of significant size and homogeneity

An outline of the general methodological approach is presented in figure 3 as an overview of how data flows from being gathered to the final regression models. The steps in this method were specifically selected to ensure that data would be handled and analyzed according to the criteria above. Each stage of the process as well as rationale for the model selection and adherence to the study's goals is further examined and elaborated on in more detail below.

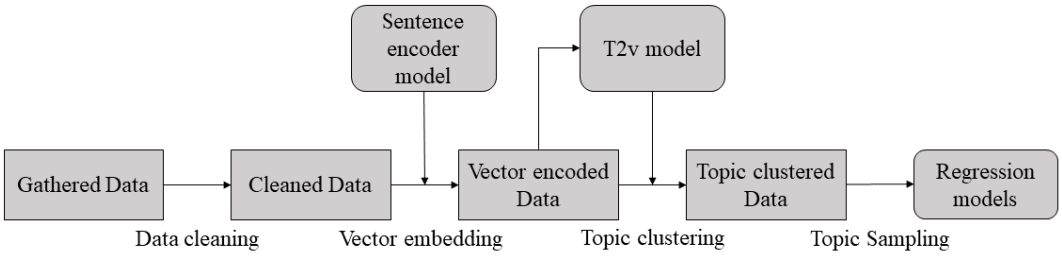

Figure 3. Outline of methodology

3.2 Data

Data was acquired from the social media platform Reddit using its API (Reddit, 2022). 2.4 billion datapoints were gathered, which included all comments and posts made between 01.06.21-31.05.22 and consisted of the posts full text, posting time, and number of likes

received. Reddit is a social media platform founded in 2005 (Stafford, 2016), that receives 1,7 billion visits per month, making it the 10th most visited website in the world (statista, 2023). All of the content on the platform is user submitted, and includes links, text, images, and videos which are then voted on in the form of likes.

Reddit was chosen as the source of data in this study for two main reasons. The first reason is the amount of available data and its ease of access, allowing the study to meet the previously outlined sample size requirement. The second reason is the fact that the platform does not allow users to follow each other, and content presented on the platform is therefore less user individualized. Engagement with the content is thus more dependent on the content itself rather than its environment such as user following and communities, which allows the study to not need to control for factors such as followers of individual accounts or number of members of a group.

Because of limited computational and storage capacity, A sample of 1% was drawn randomly from the dataset, resulting in a workable dataset of 24 million comments and posts. This dataset was further reduced to 18,965,138 datapoints by removing all posts that contained two words or less, had less than 0 likes, or contained profanity. The data was then cleaned following a method similar to Gregoriades et al., (2021) and Albalawi et al., (2020) by removing website links, moving all text to lower-case, lemmatizing, removing non-alphabetic characters, and replacing newline symbols with spaces. The method used for data cleaning was aimed both at removing useless or low-value posts, as well as simplifying the posts to achieve more generalized vector embeddings.

3.3 Vector embedding

A BERT sentence encoder (Devlin, et al., 2019) with a bigram vocabulary was used to create 384-dimensional document vector embeddings of the text of each of the sampled posts in the dataset. The BERT model was used because of the quality of its vector embeddings (Reimers, 2022), as well it being open source, allowing the method to be easily reproduced.

3.4 Short text topic modelling

A Topic2Vec model was then trained and used to cluster the posts based on their semantic topic (Niu & Dai, 2015). This model was chosen specifically for its high performance on short texts, as well as its ability to produce few “useless” topics, resulting in as many topics as possible being represented by a significant number of posts. A subsample of 1,800,000

posts was randomly drawn from the dataset for training the model as the training algorithm scales in complexity at a rate of approximately *O(nlog(n))* with respect to number of samples, which made training the model on the full dataset infeasible.

To train the Topic2Vec model, Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) was applied to the document embeddings, with number of neighbors parameter set to 50 to reduce the vector dimensionality to 5. This dimension reduction step is essential in reducing the distance between vectors, ensuring that clustering can take place. Finally, Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) was performed on the resulting vectors with a minimum cluster size set to 15, and Excess of Mass as the clustering selection method (Niu & Dai, 2015). This resulted in 11308 different topic vectors which were then scaled back up to 384 dimensions using the UMAP embeddings. The parameters used were chosen because they are the default of the Topic2Vec model as outlined in Niu & Dai (2015).

This Topic2Vec model was then used to categorize each post in the full dataset by assigning each post to a topic based on the document vectors cosine similarity to the topic vectors. Cosine similarity represents the distance between the two vectors and thus their semantic similarity. It is presented in Equation 1 (Mikolov, et al., 2013b), where A and B are the topic vector and document vector respectively.

Eq. 1 $$cosine\ similarity(A, B) = \frac{A \cdot B}{\|A\| + \|B\|}$$

3.5 Variables

As outlined in the proposed conceptual model, independent variables were chosen based on their significance in previous research, and all represent the semantics of the social media posts in one way or another.

3.5.1 Variable Extraction

3.5.1.1 Engagement

Engagement was chosen as the dependent variable in line with previous research and is represented by the number of likes that the post received, as an integer. (Lei, et al., 2017; Menon, et al., 2019; Peruta & Shields, 2018).

3.5.1.2 Semantic vector embedding

The full text of each post was encoded into 384-dimensional vectors using a BERT sentence encoder (Devlin, et al., 2019). Each of the 384 vector components were represented as individual variables as can be seen in Figure 4 and Equation 2.

Text semantics such as subjects and descriptors have previously been shown to affect engagement on social media (Peruta & Shields, 2018; Aggarwal, et al., 2009). Li & Bond (2022) have also shown how semantic similarity is related to engagement on social media, and inclusion of these variables is therefore made on the basis that semantics of subjects and descriptions can be represented using statistical word embedding models (Mikolov, et al., 2013a).

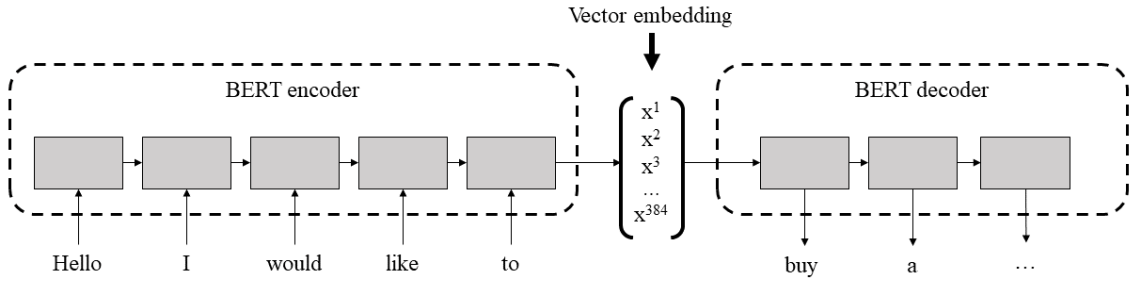

Figure 4. Visual representation of BERT encoder/decoder architecture and latent vector embedding.

3.5.1.3 Information quantity

Information quantity was represented by three different variables: number of individual words in the post, number of subjects in the post (verbs and nouns), and number of adjectives in the post.

Information quantity in a post affects a consumer's ability to process it, as well as potentially allowing for more value to be gained from the post in the form of information (de Vries, et al., 2012). This factor was included since informational appeal as well as post length have previously been proven to affect engagement (Cervellon & Galipienzo, 2015; Schultz, 2017).

Part of speech was extracted from each post using a Greedy Averaged Perceptron Tagger with the help of the open-source NLTK package (Honnibal, 2013). The part of speech information was then used to count the number of subjects and adjectives.

3.5.1.4 Uniqueness

Uniqueness is usually measured in relation to the context of a post and has been shown to have a significant effect on engagement (Pino, et al., 2019; Huertas & Marine-Roig, 2016).

This variable is included on the basis that distinctive posts might draw increased attention from consumers, and the study therefore aims to represent uniqueness in terms of a posts semantic similarity to the topic to which it has been assigned.

Uniqueness was calculated as the cosine similarity between the vector embedding of the post and the topic vector. This variable is represented by a number in the range [0, 1] where the more unique a post is the lower the cosine similarity.

3.5.1.5 Sentiment

Valence of sentiment of the text was calculated using an open-source Valence Aware Dictionary and sEntiment Reasoner (Hutto, 2022; Hutto & Gilbert, 2014). This variable is represented by a number in the range [-1.0, 1.0], where -1.0 is the most negative, and 1.0 is the most positive. Sentiment is included because of its inclusion and significance in previous studies such as (Viswanathan, et al., 2018).

3.5.1.6 Timing

Finally, control variables based on posting time was used as per previous research. This allowed the study to adjust for possible seasonal topics, or other consumer activity cycles.

Three indicator variables were created:

- Month of posting
- Weekday of posting
- Hour of posting

Constituting 40 individual variables in total.

3.6 Statistical Tests

Two statistical tests were conducted on the posts and their engagement with the aim to find out if the number of likes is normally or Poisson distributed. The results of the two statistical tests were then used to choose the regression model with the highest amount of explanatory potential. (Lei, et al., 2017)

Testing for a normal distribution was done through a D'Agostino's K-squared test for normality (D'Agostino, 1971). The test measures goodness-of fit of the data to a corresponding normal distribution with the null hypothesis that the data is sampled from an independent identically distributed gaussian variable. It works on the basis that a normal

distribution has a skew and kurtosis of 0 and therefore rejects the null hypothesis if either of these are significantly different than 0.

A Poisson distribution was tested for using a Pearson Chi-Square test based on Lei et al., (2017), with a null hypothesis that the data was sampled from a discrete Poisson distribution. As a Poisson distribution is defined as having variance equal to its mean, the test rejects the null hypothesis if the data is skewed in such a way that the mean of the data is significantly different from its variance.

3.7 Regression modelling

30 topics were randomly selected from all topics containing 5000 or more posts. This threshold was put in place to avoid problems relating to degrees of freedom as a result of the large number of parameters. Variables were then extracted from each post in each of the topics and placed in 30 different datasets according to topic.

A negative binomial regression model (Negbin II) (Cameron & Trivedi, 1986) was then defined for each topic as per Equation 2:

Eq. 2 $$ln\,(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \beta_{n+1} v_1 + \beta_{n+2} v_2 + \cdots + \beta_{n+384} v_{384}$$

Where y is the dependent variable, α is the intercept parameter, β_n are the regression coefficient parameters for each independent variable, v_n are the individual scalars from the semantic vector embedding of each post, and x_n are the other variables extracted from each post, including the adjustment variables.

Negative binomial regression is an extension of Poisson regression with an extra parameter to model over-dispersed datasets. It was chosen as the confidence interval for the model is likely narrower compared to that of a Poisson regression for the data used.

From the model, p-values and parameters were extracted for analysis. R-squared values were also calculated for each regression model using McFadden's pseudo R-squared. (McFadden, 1974)

3.8 Validity and Reliability

Text embeddings are high-dimensional, abstract, and opaque, and the validity of using document vector embeddings to model semantics has therefore received significant attention in previous studies. Fang et al., (2022) conducted a study specifically for evaluating the

validity of such embeddings with respect to information to be encoded and tested multiple models against these benchmarks. They found that different embedding models demonstrate different degrees of validity, but that BERT sentence encoders were among the best in terms of performance across all benchmarks (Fang, et al., 2022). As a BERT sentence encoder was used in this study, validity is likely not an issue.

Reliability of the method is affected by two main factors: source and sampling. Since the source of the data was a single social media channel, the applicability of this study might be limited in terms of other social media platforms. The data was also gathered during a specific time period, which also affects reliability as data from different time periods might not show the same results. This was however limited to some degree by extending the time period to be as wide as possible in order to ensure minimum impact on reliability in terms of time of year or month. As was presented in the method, data is sampled and sub-sampled multiple times as it is processed. Data was sampled randomly and uniformly to minimize the effect it would have on the method's reliability, but would still affect results should the method be repeated. The large sample sizes also ensure that the sampling has minimal effect on the reliability of the findings in the study.

4 Results

4.1 Descriptive statistics

4.1.1 Engagement

After sampling and cleaning, 18,965,138 social media posts remained in the workable dataset. Figure 4 shows the distribution of posts according to number of likes received. The average number of likes was 9.22, while the median was 2. 3.2% of posts had 0 likes, 95.0% of posts had less than 24 likes, and 1.1% of posts had more than 100 likes.

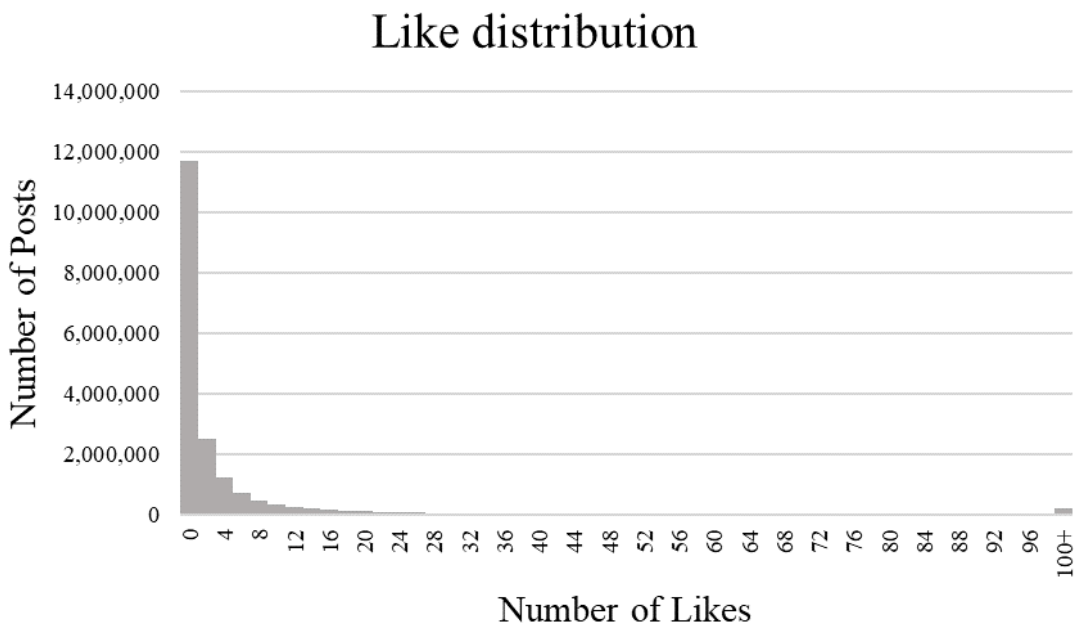

Figure 5. Distribution of engagement in the cleaned dataset

4.1.2 Topics

A distribution of the topics and number of posts assigned to each is represented in Figure 6. Each topic had on average 1,677 posts assigned to it, with a median of 938. The largest topic was assigned 88,930 posts while the smallest was assigned 2 posts.

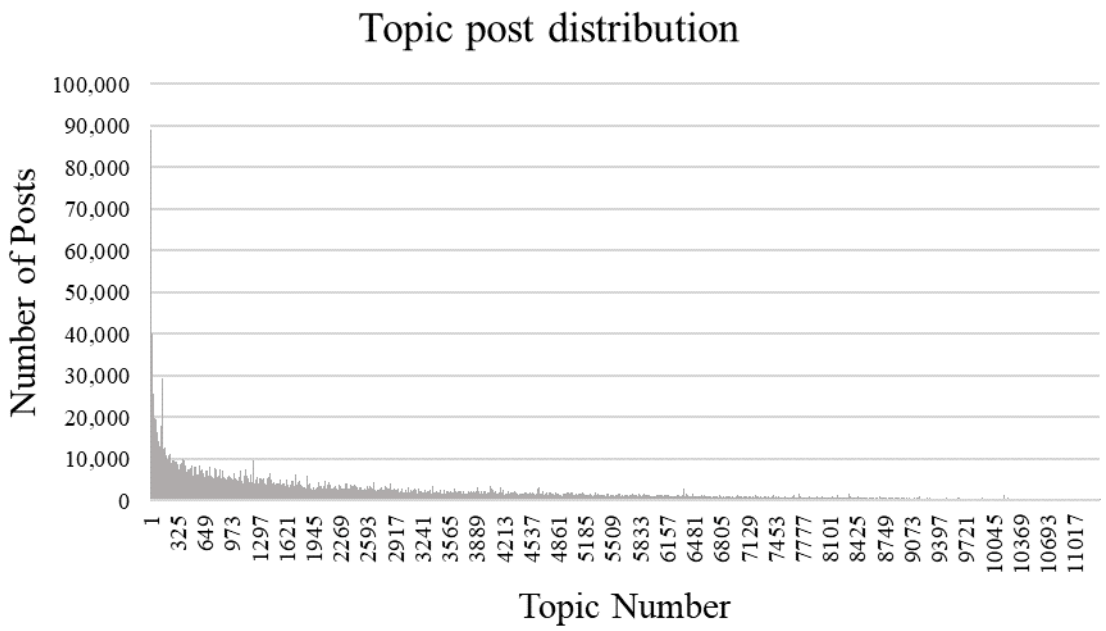

Figure 6. Number of posts assigned to each topic.

A sample of documents and assigned topics can be seen in Table 1. The words representing the topic are the five words closest to the topic vector in terms of cosine similarity, some of these are two words since the BERT sentence encoder made use of a bigram vocabulary.

Original Post	Words representing assigned topic
I think if they played twice, they'd split the two games. Can I have both? Such a hard decision, though. If you had LSU, you could win shootout after shootout. If you had Georgia, you wouldn't be in that situation. Can't say no to Jordan Davis and Nakobe Dean.	'ohio state' 'osu' 'college football' 'bama' 'boise state'
Thank God, Congress limits the Draft these days to great emergencies and war.	'combat veteran' 'veterans' 'war veteran' 'military veteran' 'veteran'
I try and tell the homies how my life is infinitely better because they're in it, the impact they've had and how much they're loved, trusted and respected. And that I'm always here for them in whatever capacity they need. No judgement, just here.	'pursue relationship' 'serious relationship' 'relationship advice' 'relationship' 'reconsider relationship'
Short answer? No. If it specifically says military plus spouse, then the military person buying the ticket has to pick up the ticket. I doubt they check the spouse thing too closely. Otherwise there would be a raging black market of discount tickets.	'buy ticket' 'ticket' 'ticket sale' 'ticketmaster' 'concert ticket'
Indeed. I have grown impatient with your circular logic and immature refusal to even consider my main point: Futurism is abusing your credulity.	'fallacy' 'logical fallacy' 'fallacious' 'logic flaw' 'flawed logic'

Table 1. Sample of social media posts and respective assigned topics.

4.2 Statistical tests

Table 2 presents the results from the D'Agostino's K-squared test for normal distribution displaying k2 statistic as well as p value. The null hypothesis was rejected as the p-value was lower than the threshold of 0.05.

	Normal distribution test
k2	17612763.9
p value	0.000

Table 2. Results from D'Agostino's K-squared test for normal distribution.

Table 3 presents the results from the Pearson Chi-Square test for sample variance and mean. The calculated Q value was larger than the critical Q for the p-value of 0.05 thus rejecting the null hypothesis.

	Pearson Chi Square test
Calculated Q	385565205.6
Critical Q	18738611.9
p value	0.000

Table 3. Results from Pearson Chi-Square test.

As both the test for normal distribution and the test for Poisson distribution rejected their null hypothesis, a negative binomial regression model was chosen to ensure the best possible estimation.

4.3 Regression

Table 4 and 5 show the results of each of the regression models in terms of what parameters were significantly different from 0 with a p-value of less than 0.05. Table 4 specifically focuses on the variables from the semantic vector embeddings, with the fourth column showing how many of the 384 vector embedding scalars were significant. Since one model was estimated for each topic, the number of significant vector-embedding variables essentially represents the strength in correlation between semantics and engagement for that topic. The fifth column in table 4 presents the calculated pseudo r-squared for each regression model which ranges from 0.09 to 0.97 with an average of 0.47.

Topic representation	Topic id	number of posts	number of significant variables (out of 384)	Psuedo R-squared
'consistent' 'consistency' 'lack consistency'	1	5396	98	0.20
'couple year' 'years' 'three month'	2	13622	137	0.12
'fascist' 'fascism' 'rise fascism'	3	6385	146	0.32
'platinum' 'gold platinum' 'platinums'	4	5911	188	0.81
'coffee' 'drinking coffee' 'cup coffee'	5	6652	129	0.38
'chinese govt' 'chinese government' 'manufacture china'	6	9475	116	0.09
'fire' 'catch fire' 'fuel fire'	7	6801	244	0.84
'infantry unit' 'infantry' 'heavy infantry'	8	7039	85	0.20
'degree' 'bachelor degree' 'college degree'	9	5141	155	0.70
'drive' 'drive safely' 'driveshaft'	10	5313	130	0.60
'wear helmet' 'helmet' 'bicycle helmet'	11	5937	195	0.47
'apple store' 'payment via' 'confirm payment'	12	5628	99	0.37
'opinion unpopular' 'controversial opinion' 'opinion invalid'	13	6431	141	0.31
'map' 'google map' 'map marker'	14	13691	176	0.33
'crypto' 'cryptos' 'crypto com'	15	5764	70	0.25
'throw ball' 'ball' 'dropping ball'	16	11817	151	0.10
'test positive' 'false positive' 'test negative'	17	7754	211	0.98
'power' 'power dynamic' 'willpower'	18	5845	216	0.76
'song' 'favorite song' 'popular song'	19	8486	226	0.50
'housing market' 'income housing' 'affordable housing'	20	10237	57	0.12
'horse' 'ride horse' 'horseback'	21	6624	223	0.83
'rain' 'weather rain' 'rain storm'	22	5511	221	0.88
'weirdness' 'weirdo' 'weirdly'	23	5002	169	0.44
'lyric' 'language lyric' 'song lyric'	24	5533	204	0.76
'smoking weed' 'smoke weed' 'weed'	25	7601	162	0.44
'bakery' 'local bakery' 'baker'	26	7848	247	0.16
'wildlife' 'local wildlife' 'habitat'	27	5735	193	0.51
'defense offense' 'offensive defensive' 'offense'	28	7140	139	0.27
'drink' 'drink alcohol' 'consume alcohol'	29	7074	225	0.96
'wet' 'water wet' 'wetness'	30	6715	203	0.31

$p < 0.05$

Table 4. Selected topics and number of significant vector-embedding variables for each

Figure 7 represents a histogram of the number of significant vector embedding variables for each topic. The average topic had 165.5 significant variables, with 247 being the highest and 57 being the lowest, and a standard deviation of 51.8. The number of significant variables was therefore very different between topics, however, the sample size might be too small to say anything meaningful about its distribution.

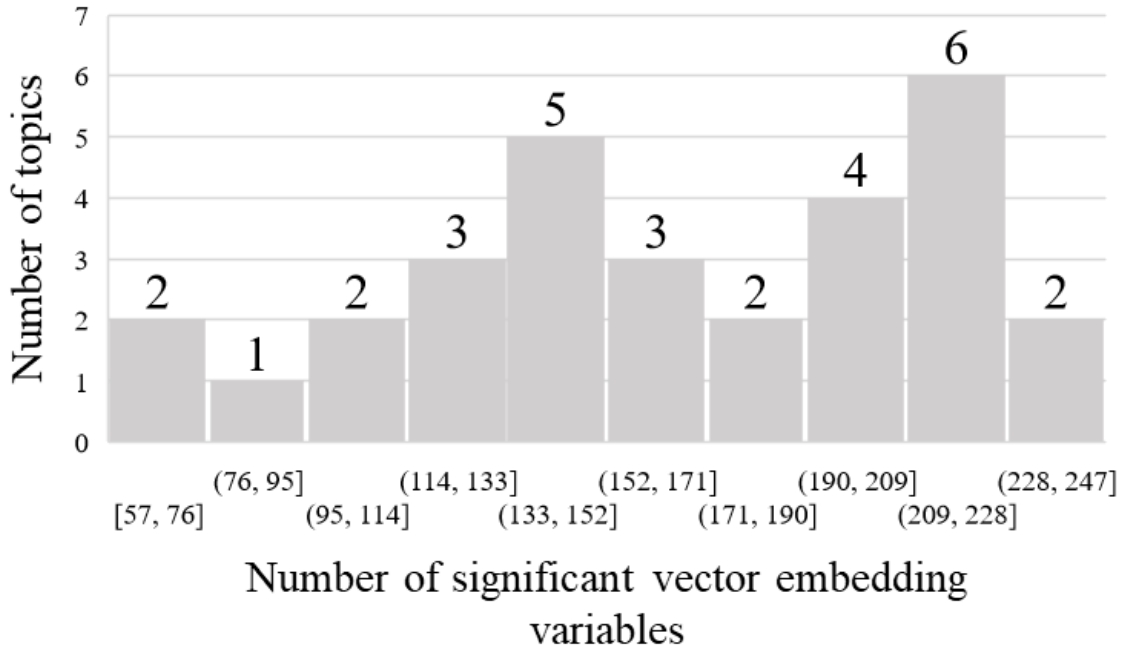

Figure 7. Histogram of number of significant vector embedding variables for topics.

In table 5, the significance of the other semantic variables used in the regression were shown, where a '1' indicates significance at a threshold of $p < 0.05$ and '-1' indicates no significance. Uniqueness was significant for 19 of the topics, while sentiment was significant for 15. The information quantity variables (length, number of subjects, number of adjectives) were significant for 16, 14, and 13 topics respectively.

Topic id	Uniqueness	Post length	Subject count	Adjective count	Sentiment
1	-1	-1	-1	*1	*1
2	*1	*1	-1	*1	*1
3	-1	-1	-1	-1	*1
4	*1	-1	-1	-1	-1
5	-1	*1	*1	*1	*1
6	*1	-1	-1	-1	-1
7	*1	*1	-1	*1	-1
8	*1	*1	*1	*1	-1
9	*1	*1	-1	-1	*1
10	-1	-1	-1	*1	*1
11	*1	-1	*1	*1	-1
12	-1	*1	-1	-1	-1
13	*1	-1	*1	-1	-1
14	-1	*1	*1	*1	-1
15	*1	-1	-1	-1	-1
16	*1	-1	-1	-1	-1
17	*1	*1	*1	-1	*1
18	*1	-1	-1	-1	-1
19	-1	-1	*1	-1	*1
20	-1	-1	*1	-1	-1
21	*1	*1	*1	-1	*1
22	*1	*1	*1	*1	*1
23	-1	-1	-1	-1	-1
24	-1	-1	-1	*1	*1
25	*1	*1	*1	-1	-1
26	-1	*1	*1	-1	*1
27	*1	*1	-1	*1	*1
28	*1	*1	-1	*1	-1
29	*1	*1	*1	-1	*1
30	*1	*1	*1	*1	*1

$* p < 0.05$

Table 5. Table indicating significance of semantic variables in regression for each topic.

Table 6 is a covariance matrix for the significance of the semantic variables, with higher values indicating that the variables tend to be significant together. Although none of the covariances between significance in the variables were particularly strong, some seem to display a relationship to the others such as length and number of subjects.

	Uniqueness	Length	Subject count	Adjective count	Sentiment
Uniqueness	0.93	0.25	0.02	-0.03	-0.20
Length	0.25	1.00	0.34	0.28	0.27
Subject count	0.02	0.34	1.00	-0.01	0.13
Adjective count	-0.03	0.28	-0.01	0.98	0.20
Sentiment	-0.20	0.27	0.13	0.20	1.00

Table 6. Covariance matrix for significance in semantic variables.

5 Discussion

5.1.1 Document vector embeddings for modelling engagement

As presented, the semantic vector embeddings of the text in social media posts show significance in each of the 30 topics that regression models were estimated on. This degree of significance combined with the r-squared values gives some credence to the idea that the semantic vector embeddings hold exploratory potential with regard to modeling engagement with the content from which they were created. Since the embeddings are previously proven to represent semantic information, this leads to the conclusion that semantic information in social media posts has a measurable effect on the engagement of that post, and that language embedding models are a viable way of quantitatively modeling this semantic information.

As engagement in social media posts is indicative of consumer behavior and perception (Brodie, et al., 2011), this is essentially a quantitative way of measuring the impact of the semantics within text on that consumer behavior. This expands on the theories in Aggarwal et al., (2009) and Hatzivassiloglou & McKeown (1997) for how semantic factors affect consumer preception of messages, brands, and content, and specifically highlights that content semantics contains information that can inform marketing decision-making, and that these semantic factors can be modeled.

In practical terms, the method presented allows for businesses and academics alike to begin modeling semantics in a quantitative way, which allows them to measure the impact that the language they use has on the engagement that their content receives and potentially start to improve the way that they communicate. This could be particularly helpful in the research vein focusing on two-way marketing communication between consumers and businesses (Aydinn, 2019). It could also allow for further solutions to be developed down the line with respect to better understanding of the content semantics and how they can be changed in practical ways to better engage consumers. In this sense, the results of this study could lead to more data-driven decision making as well as a better understanding of consumer interaction with marketing content, as has been called for by previous research (Kujur & Singh, 2018; Alves, et al., 2016; Aydinn, 2019).

The results regarding content semantics could also have an influence on the understanding on cognitive social identification relating to SIT. Further implications regarding group identification and belonging could be made by modeling the semantic differences between

language used in different groups. This would enable researchers and marketers to better understand how consumers identify with groups and how they display that identification, allowing for improved brand community engagement as well as how interactions differ in separate groups on social media (Kumar & Kumar, 2019; Kumar & Nayak, 2018). This has the potential to improve both brand loyalty as well as increase customer evangelization through electronic word of mouth (Alalwan, et al., 2017).

5.1.2 Topics as a moderating variable

As presented, the degree of significance of the vector embeddings varies between topics, with some having close to five times more significant vector embedding variables. This implies that the topic discussed to some degree moderates the relationship between the semantic variables and the resulting engagement. For the previously established correlation between content semantics and the consumer behavior, the topic of the content is indicative not only of how the content is perceived, but also how important the semantic information in the content is. This means that marketers must adapt their marketing strategies depending on this topic context and might be required to adapt the way they communicate for different communities.

Although not accounted for in this study, there are many possible causes of the moderating effect that topics have topics such as the demographics of the users engaging with different topics, how the content is consumed, or different topics being discussed in different communities. The relationship that needs to be better understood in order to establish this, is whether the topic directly affects the consumer behavior, or if it simply is a matter of different demographics engaging with different topics. The study by Peruta & Shields (2018) covers engagement factors in different topics on a small scale, but this study emphasizes that larger scale studies would be required in order to establish the relationship between topic and consumer behavior factors.

The differing degree of significance in semantics for each of the topics could give marketers a clue on what to prioritize for the content they are trying to create. If, for instance, semantics have a low degree of significance for a particular topic the marketer can divert resources to other facets of their marketing strategy. This would help marketers become more effective while still allowing for the best possible quality content to be generated.

As shown in Aggarwal et al., (2009) language on social media can have a significant impact on consumer relationships with brands and the brands positioning in the market. The importance of topics presented in the results of this study could be a further avenue of study with regard to brand positioning and perception, as topic might also moderate these relationships. Findings in this research vein could improve long-term marketing strategies such as market positioning, as well as reduce the need for focus groups or surveys and thus increase industry efficiency.

The study's findings on topics also suggest that both marketers and academics should place significant attention on the bias that may form because of the source of data. Marketers specifically might have to adapt their marketing techniques depending not only on who they are communicating with but what they are trying to communicate about, as this might affect which factors are significant with regard to creating engagement.

5.1.3 Complimentary variables

The other variables modeled: uniqueness, sentiment, post length, number of subjects, and number of verbs, all show significance in many of the regression models estimated.

Uniqueness is particularly interesting for this study as it was also calculated using the cosine distance between vector embeddings, which lends further credibility to using the method to model semantic information. Although the relationship between uniqueness and engagement has been studied before (Pino, et al., 2019), this study offers some novelty in measuring semantic uniqueness quantitatively as opposed to with human labelling, and therefore further elaborates on the significance of semantics in social media content on consumer behavior. As the method is machine-based it also eliminates bias in the data-generation process that might arise from human labelling, allowing for future studies that want to examine uniqueness to become more robust and replicable.

The significance of uniqueness might also present a shift in perspective regarding research on trends. The ability to measure the effectiveness of alternate strategies, such as going against trends, could allow marketers to be more creative in the content creation process as well as their strategic decision-making. As the amount of content on social media increases, using information about uniqueness in relation to consumer behavior could also increase marketers ability to create content that stands out and attract users' attention, reducing the growing risk of content getting lost in the noise (Aydinn, 2019; Xu, et al., 2014).

Sentiment has been used as a variable in many previous studies such as Viswanathan et al., (2018), but its significance in this study come with two main implications. The first is that since it has been proven significant in previous studies, it improves the credibility of the findings in this paper, as not only was it found significant itself, but several other variables such as uniqueness and semantics were found to be significant even more often. The second is that its presense does not reduce the informational value of the other variables as can be seen in the covariance matrix, indicating that what is being modelled by the semantic vectors is not sentiment, and that using vector embeddings therefore add a significant degree of informational value.

Of the three variables representing information quantity, post length was the one with significance for the largest number of topics. However, as can be seen in the significance covariance matrix, the other variables each provide further informational value to different topics, suggesting that they should be used together to achieve the highest possible goodness-of-fit to an unknown dataset. Post length is a variable that has seen extensive research, but the significance of other variables that represent information quantity show that there is still potential research to be made on the subject (Schultz, 2017).

5.1.4 General implications

The first goal of this study was to assess the potential of using vector embeddings to model engagement. I consider the results of the study to have achieved this goal by showing some degree of significance in the vector embedding variables for all topics. The second goal of the study was to gauge the moderating effect topics would have on the relationships modeled. Since it was shown that the variables that are significant change considerably between topics, I also consider this goal to have been met.

The method presented could help solve some of the problems in the academic field of content marketing. More specifically, the difference in significant factors between topics might play a role in the number of contradictory findings in previous studies (Gutiérrez-Cillán, et al., 2017; Schultz, 2017). Taking topic differences into account as a moderating variable could therefore lead to better understanding of other consumer behavior factors as well as their differences between communities.

The method’s reliance on automatic and quantitative processees is also something that could benefit further academic studies. The method allows for large amounts of data to be

processed as it avoids human-labelling bottlenecks, and might also therefore also help future studies with regards to increasing their sample sizes, and thus significance of their findings.

In his paper, Aydinn (2019) outlines four facets that affect post engagement on social media: technology acceptance, user motives and values, psychological factors and attitudes, and post characteristics. The method in this study highlights a path forward for better understanding the last two facets by extending the toolbox of potential factors academics can use when modelling engagement in different scenarios. This could allow them to both improve the explanatory potential in their models, adjust for variables that were previously unaccounted for, as well as expand the reach of potential research focus.

The approach essentially contributes to the vein of research that focuses on understanding consumers better with the information freely available to businesses on the web. This allows them not only to better understand their own customers but also their position in the market on an international scale. As the information available changes over time, this field of research is also key to measure changes in the market, promoting adaptation and innovation.

5.2 Future Research

Although the study displays significance in the explanatory potential of text vector embeddings regarding engagement, it does not try to derive any actionable suggestions for content marketing. Future research could make this impact by performing a more detailed study, considering the valence of estimated parameters for specific topics in order to create a better understanding of the best course of action for marketers within that topic. This could be further amplified by studies that aim to improve the explanatory potential of vector embeddings, which would help create actionable insights from the vector embedding parameters.

Another potential path for future research is development of the method using different embedding models that could either increase the quality of the word embeddings or improve their interpretability. Advances along this path could see an increase in the applicability of vector embedding models to real life marketing problems. Further research into different methods for clustering the vectors could also be of benefit, as these might result in better differentiated topics or topics of different kinds.

Lastly, relating the method presented back to the fundamental CM theories: Uses and Gratification Theory and Social Identity Theory, would contextualize possible findings better.

This would give us a more nuanced understanding of why content semantics affect engagement in the way it does and provide further insights into how to adapt marketing efforts to this.

6 Conclusion

Developments in business analytics as well as an increased availability of data has allowed digital marketers to better understand and capitalize on consumer behavior to maximize the engagement with marketing materials (Lies, 2019). Because most previous studies in this field have focused on consumer behavior theory, they have been largely limited in scope due to small datasets and reliance on human-labeled data.

This study aimed to explore the potential of using a machine-learning language model to generate vector embeddings, representing the semantics in text, to model engagement in a quantitative way. By clustering the semantic vector embeddings, the study generated datasets on different topics, on which regression models were estimated to gauge the impact of the represented variables. Many of the parameters in the models were shown to be significant, implying both explanatory potential in text semantics, as well as the presented methods ability to model these.

This study expands on theories in the literature regarding how semantic factors affect consumer perception, as well as highlighting that textual semantics contains information that can help inform marketing decision-making. Its main contribution is a methodology that can allow academics and marketers alike to model these semantics and thus gain insights relating to how topics and language affect consumer engagement. Further investigation into similar methods might allow digital marketers to improve their understanding how different consumers perceive and engage with their marketing content.

7 References

Aggarwal, P., Vaidyanathan, R. & Venkatesh, A., 2009. Using Lexical Semantic Analysis to Derive Online Brand Positions: An Application to Retail Marketing Research. *Journal of Retailing,* 85(2), pp. 145-158.

Alalwan, A. A., Rana, N. P., Dwivedi, Y. K. & Algharabat, R., 2017. Social Media in Marketing: A Review and Analysis of the Existing Literature. *Telematics and Informatics,* 34(7), pp. 1177-1190.

Albalawi, R., Yeap, T. H. & Benyoucef, M., 2020. Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. *Frontiers in Artificial Intelligence,* 3(42).

Alves, H., Fernandes, C. & Raposo, M., 2016. Social Media Marketing: A Literature Review and Implications. *Psychology & Marketing*, December, pp. 1029-1038.

Asmussen, B. et al., 2015. How do you conceptualise branded content? An exploration of marketing communications discourse. *Academy of Marketing, Limmerick,* 9 July.

Aydinn, G., 2019. Social media engagement and organic post effectiveness: A roadmap for increasing the effectiveness of social media use in hospitality industry. *Journal of Hospitality Marketing & Management,* 29(1), pp. 1-21.

Baldus, B. J., Voorhees, C. & Calantone, R., 2015. Online brand community engagement: Scale development and validation.. *Journal of Business Research,* 68(5), pp. 978-985.

Beukeboom, C. J., Kerkhof, P. & Vries, M., 2015. Does a Virtual Like Cause Actual Liking? How Following a Brand's Facebook Updates Enhances Brand Evaluations and Purchase Intention. *Journal of Interactive Marketing,* 32(1), pp. 26-36.

Bianchi, C. & Andrews, L., 2018. Consumer engagement with retail firms through social media: An empirical study in Chile. *International Journal of Retail and Distribution Management,* 46(4), pp. 364-385.

Blumler, J. G., 1979. The role of theory in uses and gratifications studies. *Communication Research,* 6(1), pp. 29-36.

Brodie, R., Hollebeek, L. D., Juric, B. & Ilic, A., 2011. Customer engagement: Conceptual domain, fundamental propositions, and implications for research. *Journal of Service Research,* 17(3), pp. 1-20.

Brodie, R., Ilic, A., Juric, B. & Hollebeek, L., 2013. Consumer engagement in a virtual brand community: an exploratory analysis. *Journal of Business Research,* 66(1), pp. 105-114.

Brown, L. D. & Zhao, L. H., 2002. A Test for the Poisson Distribution. *The Indian Journal of Statistics,* 64(3), pp. 611-625.

Bruce, R. F. & Wiebe, J. M., 1999. Recognizing Subjectivity: A Case Study of Manual Tagging. *Natural Language Engineering,* Volume 5, pp. 187-205.

Cameron, A. C. & Trivedi, P. K., 1986. Econometric models based on count data. Comparisons and applications of some estimators and tests.. *Journal of Applied Econometrics,* 1(1), p. 29–53.

Cervellon, M. C. & Galipienzo, D., 2015. Facebook pages content, does it really matter? consumers' responses to luxury hotel posts with emotional and informational content. *Journal of Travel and Tourism Marketing,* 32(4), pp. 428-437.

Chen, Y., Peng, Z., Kim, S.-H. & Choi, C. W., 2023. What We Can Do and Cannot Do with Topic Modeling: A Systematic Review. *Communication Methods and Measures,* 17(2), pp. 111-130.

Cheung, C., Zheng, X. & Lee, M., 2012. *Consumer engagement behaviors in brand communities of social networking sites.* Seattle, Americas Conference on Information Systems.

Chiche, A. & Yitagesu, B., 2022. Part of speech tagging: a systematic review of deep learning and machine learning approaches. *Journal of Big Data,* 9(10).

Chi, H.-H., 2011. Interactive Digital Advertising VS. Virtual Brand Community: Exploratory Study of User Motivation and Social Media Marketing Responses in Taiwan. *Journal of Interactive Advertising,* 12(1), pp. 44-61.

Coelho, R. L. F., de Oliveira, D. S. & de Almeida, M. I. S., 2016. Does social media matter for post typology? Impact of post content on Facebook and Instagram metrics. *Online Information Review,* 40(4), pp. 458-471.

Cvijikj, I. P. & Michahelles, F., 2013. Online engagement factors on Facebook brand pages. *Social Network Analysis and Mining,* 3(4), pp. 843-861.

D'Agostino, R. B., 1971. An Omnibus Test of Normality for Moderate and Large Size Samples. *Biometrika,* 58(2), pp. 341-348.

de Clerck, J., 2013. *This is your real social business strategy challenge.* [Online] Available at: https://www.i-scoop.eu/real-social-business-strategy-challenge/ [Accessed 21 April 2023].

de Oliveira Santini, F. et al., 2020. Customer engagement in social media: a framework and meta-analysis. *Journal of the Academy of Marketing Science,* 48(1), p. 1211–1228.

de Vries, L., Gensler, S. & Leeflang, P. S. H., 2012. Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing. *Journal of Interactive Marketing,* 26(2), pp. 83-91.

Dessart, L., loutsou, C. & Morgan-Thomas, A., 2016. Capturing consumer engagement: duality, dimensionality and measurement. *Journal of Marketing Management,* pp. 1-28.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT, Minneapolis,* 2-7 June.

Dolan, R., Conduit, J., Fahy, J. & Goodman, S., 2016. Social media engagement behaviour: A uses and gratifications perspective.. *Journal of Strategic Marketing,* 24(3), pp. 261-277.

Dwivedi, Y. K. et al., 2021. Setting the future of digital and social media marketing research: Perspectives and research propositions. *International Journal of Information Management,* 59(1).

Dwivedi, Y. K., Kapoor, K. K. & Chen, H., 2015. Social media marketing and advertising. *The marketing review,* 15(3), pp. 289-309.

Fang, Q., Nguyen, D. & Oberski, D. L., 2022. Evaluating the construct validity of text embeddings with application to survey questions. *EPJ Data Science*, pp. 11-39.

Farook, F. S. & Abeysekara, N., 2016. Influence of Social Media Marketing on Customer Engagement. *International Journal of Business and Management Invention,* 5(12), pp. 115-125.

Felix, R., Rauschnabel, P. & Hinsch, C., 2017. Elements of strategic social media marketing: a holistic framework. *Journal of Business Research,* 70(1), pp. 118-126.

Godey, B. et al., 2016. Social media marketing efforts of luxury brands: influence on brand equity and consumer behavior. *Journal of Business Research,* 69(12), pp. 5833-5841.

Gregoriades, A., Pampaka, M., Herodotou, H. & Christodoulou, E., 2021. Supporting digital content marketing and messaging through topic modelling and decision trees. *Expert Systems With Applications,* 184(1).

Guesalaga, R., 2015. The use of social media in sales: Individual and organizational antecedents, and the role of customer engagement in social media. *Industrial Marketing Management,* 54(1), pp. 71-79.

Gummerus, J., Liljander, V., Weman, E. & Pihlström, M., 2012. Customer engagement in a Facebook brand community. *Management Research Review,* 35(9), pp. 857-877.

Gutiérrez-Cillán, J., Camarero-Izquierdo, C. & San José-Cabezudo, R., 2017. How brand post content contributes to user's Facebook brand-page engagement. The experiential route of active participation. *BRQ Business Research Quarterly,* 20(4), pp. 258-274.

Hamari, J., Koivisto, J. & Sarsa, H., 2014. Does Gamification Work? — A Literature Review of Empirical Studies on Gamification. *International Conference on System Science, Hawaii,* 6-9 Jan.

Hamilton, M., Kaltcheva, V. D. & Rohm, A. J., 2016. Social Media and Value Creation: The Role of Interaction Satisfaction and Interaction Immersion. *Journal of Interactive Marketing,* 36(1), pp. 121-133.

Harrando, I., Lisena, P. & Troncy, R., 2021. Apples to Apples: A Systematic Evaluation of Topic Models. *Proceedings of Recent Advances in Natural Language Processing*, 1-3 September.

Hatzivassiloglou, V. & McKeown, K. R., 1997. Predicting the Semantic Orientation of Adjectives. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 7 July.

Hatzivassiloglou, V. & Wiebe, J. M., 2000. Effects of Adjective Orientation and Gradability on Sentence Subjectivity. *The 18th International Conference on Computational Linguistics*, 31 July.

Holliman, G. & Rowley, J., 2014. Business to business digital content marketing: Marketers' perceptions of best practice. *Journal of Research in Interactive Marketing*, 8(4), pp. 269-293.

Holmes, R., 2015. *5 trends that will change how companies use social media in 2016.* [Online]
Available at: https://www.fastcompany.com/3054347/the-future-of-work/5-trends-that-will-change-how-companies-use-social-media-in-2016
[Accessed 5 October 2022].

Honnibal, M., 2013. *A Good Part-of-Speech Tagger in about 200 Lines of Python.* [Online]
Available at: https://explosion.ai/blog/part-of-speech-pos-tagger-in-python
[Accessed 01 2023].

Huertas, A. & Marine-Roig, E., 2016. Differential destination content communication strategies through multiple social media. In: *Information and Communication Technologies in Tourism.* New York: Springer, pp. 239-252.

Hughes, C., Swaminathan, V. & Brooks, G., 2019. Driving Brand Engagement Through Online Social Influencers: An Empirical Investigation of Sponsored Blogging Campaigns. *Journal of Marketing*, 83(5), pp. 78-96.

Hutto, C., 2022. *VADER-Sentiment-Analysis.* [Online]
Available at: https://github.com/cjhutto/vaderSentiment
[Accessed 1 2023].

Hutto, C. & Gilbert, E., 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media, Ann Arbor,* 1-4 June.

Ibrahim, N. F., Wang, X. & Bourne, H., 2017. Exploring the effect of user engagement in online brand communities: Evidence from Twitter. *Computers in Human Behavior,* 72(1), pp. 321-338.

Kaplan, A. M. & Haenlein, M., 2010. Users of the World, Unite! The Challenges and Opportunities of Social Media. *Business Horizons,* 53(1), pp. 59-68.

Katz, E., Blumler, J. G. & Gurevitch, M., 1973. Uses and Gratifications Research. *Public Opinion Quarterly,* 37(4), p. 509–523.

Kietzmann, J. H., Hermkens, K., McCarthy, I. P. & Silvestre, B. S., 2011. Social media? Get serious! Understanding the functional building blocks of social media.. *Business Horizons,* 54(3), pp. 241-251.

Kim, S., Koh, Y., Cha, J. & and Lee, S., 2015. Effects of social media on firm value for US restaurant companies. *International Journal of Hospitality Management,* 49(1), pp. 40-46.

Kim, W. H. & Chae, B. K., 2018. Understanding the relationship among resources, social media use and hotel performance. *International Journal of Contemporary Hospitality Management,* 30(9), pp. 2888-2907.

Korfiatis, N., Stamolampros, P., Kourouthanassis, P. & Sagiadinos, V., 2018. easuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews. *Expert Systems with Applications Fortcoming,* 116(1), pp. 472-486.

Kujur, F. & Singh, S., 2018. Antecedents of relationship between customer and organization developed through social networking sites. *Management Research Review,* 42(1), pp. 2-24.

Kumar, J. & Kumar, V., 2019. Drivers of brand community engagement. *Journal of Retailing and Consumer Services,* 54(1).

Kumar, J. & Nayak, J., 2018. Brand community relationships transitioning into brand relationships: Mediating and moderating mechanism. *Journal of Retailing and Consumer Services,* 45(1), pp. 64-73.

Kumar, V. et al., 2010. Undervalued or Overvalued Customers: Capturing Total Customer Engagement Value. *Journal of Service Research,* 13(3), pp. 297-310.

Kumar, V., Bhaskaran, V., Mirchandani, R. & Shah, M., 2013. Creating a Measurable Social Media Marketing Strategy: Increasing the Value and ROI of Intangibles and Tangibles for Hokey Poke. *Marketing Science,* 32(2), pp. 194-212.

Kumar, V., Choi, J. B. & Greene, M., 2017. Synergistic effects of social media and traditional marketing on brand sales: capturing the time-varying effects. *Journal of the Academy of Marketing Science,* 45(1), pp. 268-288.

Kumar, V. & Kaushik, A. k., 2018. Building consumer–brand relationships through brand experience and brand identification. *Journal of Strategic Marketing,* 28(1), pp. 39-59.

Leeflang, P. S., Verhoef, P. C., Dahlström, P. & Freundt, T., 2014. Challenges and solutions for marketing in a digital era. *European Management Journal,* 32(1), pp. 1-12.

Lee, U., Kim, Y. J., Lim, Y. S. & Kim, M., 2015. Trait reactance moderates Facebook users' irritation with brand communication. *Social Behavior & Personality,* 43(1), pp. 829-844.

Lei, S. S. I., Pratt, S. & Wang, D., 2017. Factors influencing customer engagement with branded content in the social network sites of integrated resorts. *Asia Pacific Journal of Tourism Research,* 22(3), pp. 316-328.

Leung, X. Y. & Tanford, S., 2015. What Drives Facebook Fans to "Like" Hotel Pages: A Comparison of Three Competing Models. *Journal of Hospitality Marketing & Management,* 25(3), pp. 314-345.

Lies, J., 2019. Marketing Intelligence and Big Data: Digital Marketing Techniques on their Way to Becoming Social Engineering Techniques in Marketing. *International Journal of Interactive Multimedia and Artificial Intelligence,* 5(5), pp. 134-144.

Lin, S., Yang, S. M. M. & Huang, J., 2018. Value co-creation on social media: examining the relationship between brand engagement and display advertising effectiveness for Chinese hotels. *International Journal of Contemporary Hospitality Management,* 30(4), pp. 2153-2174.

Li, Y. & Bond, R. M., 2022. Examining semantic (dis)similarity in news through news organizations' ideological similarity, similarity in truthfulness, and public engagement on social media: a network approach. *Human Communication Research,* 49(1), pp. 47-60.

Luo, J. M., Vu, H. Q., Li, G. & Law, R., 2020. Topic modelling for theme park online reviews: analysis of Disneyland. *Journal of Travel & Tourism Marketing,* 37(2), pp. 272-285.

Marketo, 2010. *The definite guide to B2B on social media,* San Mateo: Marketo Inc..

McFadden, D., 1974. Conditional Logit Analysis of Qualitative Choice Behavior.. In: *Zarembka, P., Frontiers in Econometrics* . s.l.:Academic Press Inc , pp. 105-142.

Menon, D., 2022. Factors influencing Instagram Reels usage behaviours: An examination of motives, contextual age and narcissism. *Telematics and Informatics Reports,* 5(1).

Menon, R. V. et al., 2019. How to grow brand post engagement on Facebook and Twitter for airlines? An empirical investigation of design and content factors⋆. *Journal of Air Transport Management,* 79(1).

Mikolov, T., Chen, K., Corrado, G. & Dean, J., 2013b. Efficient Estimation of Word Representations in Vector Space. *International Conference on Learning Representations,* 2-4 May.

Mikolov, T. & Le, Q., 2014. Distributed Representations of Sentences and Documents. *31st International Conference on International Conference on Machine Learning,* 21-26 June.

Mikolov, T., Yih, W.-t. & Zweig, G., 2013a. Linguistic Regularities in Continuous Space Word Representations. *Association for Computational Linguistics: Human Language Technologies,* 10-12 June.

Niu, L.-Q. & Dai, X.-Y., 2015. *Topic2Vec: Learning Distributed Representations of Topics.* Suzhou, International Conference on Asian Language Processing.

Paquette, H., 2013. *Social Media as a Marketing Tool: A Literature Review.* [Online] Available at: https://digitalcommons.uri.edu/tmd_major_papers/

Parent, M., Plangger, K. & Bal, A., 2011. The new WTP: Willingness to participate. *Business Horizons,* 54(3), pp. 219-229.

Paruthi, M. & Kaur, H., 2017. Scale development and validation for measuring online engagement. *Journal of Internet Commerce,* 16(2), pp. 127-147.

Peruta, A. & Shields, A. B., 2018. Marketing your university on social media: a content analysis of Facebook post types and formats. *Journal of Marketing for Higher Education,* 28(2), pp. 175-191.

Pino, G. et al., 2019. A methodological framework to assess social media strategies of event and destination manage ment organizations. *Journal of Hospitality Marketing & Management,* 28(2), pp. 189-216.

Qiang, J. et al., 2020. Short Text Topic Modeling Techniques, Applications, and Performance: A Survey. *IEEE Transactions on Knowledge and Data Engineering,* 34(3), pp. 1427-1445.

Rathore, A. K., Ilavarasan, P. V. & Dwivedi, Y., 2016. Social media content and product co-creation: an emerging paradigm. *Journal of Enterprise Information Management,* 29(1), pp. 7-18.

Reddit, 2022. *api-documentation.* [Online]
Available at: https://www.reddit.com/dev/api
[Accessed 22 10 2022].

Reimers, N., 2022. *sbert.net.* [Online]
Available at: https://www.sbert.net/docs/pretrained_models.html
[Accessed 02 05 2023].

Rumelhart, D. E., Hinton, G. E. & Williams, R. J., 1986. Learning internal representations by backpropagating errors. *Nature,* 323(1), pp. 533-536.

Sahlgren, M., 2008. The distributional hypothesis. *Rivista di Linguistica,* 20(1), pp. 33-53.

Santos, Z. R., Cheung, C. M. K., Coelho, P. S. & Rita, P., 2021. Consumer engagement in social media brand communities: A literature review. *International Journal of Information Management.*

Saura, J. R., 2021. Using Data Sciences in Digital Marketing: Framework, methods, and performance metrics. *Journal of Knowledge and Innovation,* 6(1), pp. 92-102.

Schultz, C. D., 2016. Insights from consumer interactions on a social networking site: Findings from six apparel retail brands. *Electronic Markets,* 26(1), pp. 203-217.

Schultz, C. D., 2017. Proposing to your fans: Which brand post characteristics drive consumer engagement activities on social media brand pages?. *Electronic Commerce Research and Applications,* 26(1), pp. 23-34.

Schutte, T. F., 1969. The Semantics of Branding. *Journal of Marketing,* 33(1), pp. 5-11.

Stafford, C., 2016. *TechTarget.* [Online]
Available at: https://www.techtarget.com/searchcio/definition/Reddit
[Accessed 02 05 2023].

statista, 2023. *Worldwide visits to Reddit.com from December 2021 to May 2022.* [Online]
Available at: https://www.statista.com/statistics/443332/reddit-monthly-visitors/
[Accessed 02 05 2023].

Tajfel, H., 1978. *Differentiation between social groups: Studies in the social psychology of intergroup relations.* London: Academic Press.

Thongmak, M., 2015. Engaging Facebook Users in Brand Pages: Different Posts of Marketing-Mix Information. *International Conference on Business Information Systems,* 16 June.

Toti, D. et al., 2020. Detection of Student Engagement in e-Learning Systems Based on Semantic Analysis and Machine Learning. *International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, 28-30 October.

Venkatesh, S. et al., 2011. Innovations in Shopper Marketing: Current Insights and Future Research Issues. *Journal of Retailing,* 87(1), pp. 29-42.

Viswanathan, V. et al., 2018. Dynamics between social media engagement, firm-generated content, and live, and time-shifted TV viewing. *Journal of Service Management,* 29(3), pp. 378-398.

Vrontis, D., Makrides, A., Christofi, M. & Thrassou, A., 2021. Social media influencer marketing: A systematic review, integrative framework and future research agenda. *International Journal of Consumer Studies,* 45(4), pp. 617-644.

Wallace, E., Buil, I. & de Chernatony, L., 2014. Consumer engagement with self-expressive brands: brand love and WOM outcomes. *Journal of Product & Brand Management*, 23(1), pp. 33-42.

Wang, W. et al., 2022. Intent Mining: A Social and Semantic Enhanced Topic Model for Operation-Friendly Digital Marketing. *International Conference on Data Engineering*, 9 May.

Wedel, M. & Kannan, P., 2016. Marketing Analytics for Data-Rich Environments. *Journal of Marketing,* 80(6), pp. 97-121.

Xu, Y. C., Yang, Y., Cheng, Z. & Lim, J., 2014. Retaining and attracting users in social networking services: An empirical investigation of cyber migration.. *Journal of Strategic Information Systems,* 23(3), pp. 239-253.

Zhang, T., Lu, C., Torres, E. & Chen, P. J., 2018. Engaging customers in value co-creation or co-destruction online. *Journal of Services Marketing,* 32(1), pp. 57-69.

Zheng, X., Cheung, C. M. K., Lee, M. K. & Liang, L., 2015. Building brand loyalty through user engagement in online brand communities in social networking sites. *Information Technology & People,* 28(1), pp. 90-106.