MSc Management of Governance Networks
Erasmus University Rotterdam

Master thesis

Local Government City Branding on Social Media and Electronic Word-of-Mouth Communication by Tourists

A case study of Copenhagen and Tallinn

Author	Carolin Ilves
Student number	427602
Coach	dr. ir. Jasper Eshuis
Second reader	prof. dr. Joop F. M. Koppenjan
Word count	20,266
Date	10 August 2016

Summary

Place branding is gaining momentum all over the globe intriguing researchers to investigate marketing strategies of public organizations. The aim is to attract target groups, mainly residents, tourists or companies, and to enhance user engagement with the brand. Since a brand cannot be controlled in a top-down manner, it is important to invite users to partake in the process of branding. In an era where the Internet and social media have become ubiquitous the sphere of electronic word-of-mouth (eWOM) has become an interesting field to research. The central research question of this study is how local government city branding influences eWOM by tourists. The research is based on a large N, population 5,299 posts, case study comparing Tallinn, Estonia, and Copenhagen, Denmark.

The results of this study show that there is a significant relationship between the type of post, length of post and time of post on the amount of likes, comments and/or shares that the post receives. This confirms the hypotheses of this study that expected a relationship between the abovementioned factors. In a separate analysis, the tone and intensity of posts are shown to affect likes, comments or shares as well. However, some hypotheses regarding the time of post, including year, month, weekday, and hour, were only partially confirmed. This study offers a possibility for brand managers to consider the factors that influence user reactions and design effective marketing strategies that induce an active response from the target group.

Contents

List of tables

List of figures

1 Introduction

1.1 Background

City branding is gaining momentum all over the globe. Metropolitan cities such as New York (*I love New York*) and Berlin (*Be Berlin*) position themselves at the center of attention for tourists and create strong brands through diverse brand elements, such as slogans, promotions, events and recognizable symbols. City branding is not limited to the large world players, however. Smaller cities such as Tartu (*City of Good Thoughts*) and Groningen (*Nothing goes above Groningen*) follow a similar strategy, albeit at a smaller scale. It is evident that city branding has become a significant observable reality. One might know a city from a logo, symbol or some other visual characteristic without ever having been there. Branding of cities emphasizes specific attributes of a geographical location to add meaning to, for instance cities or countries. The aim is often to appeal to local citizens, tourists, or investors (Eshuis & Klijn, 2012; Eshuis, Klijn, & Braun, 2014). This study particularly focuses on tourists as a target group for city branding.

Continuous growth in Internet access has placed importance on using social media for branding. Web-based social media have the capacity to induce online collaboration and content sharing (Bryer & Zavattaro, 2011). In recent times of competing media and increasingly hasty daily lives, Facebook, Twitter and other interactive platforms have become important branding tools (Yan, 2011). This development allows a multiplicity of tourists to react to city branding by electronic word-of-mouth (eWOM) communication on social media. The Internet has the capacity to allow various participants, in addition to the brand manager, to partake in branding activities. Thus, the ample communication of, for instance tourists and residents on social media about a city contribute to an existing image and reputation of that city. This creates a reality where brands become dynamic and cannot be fully controlled in a top-down way. Users can be engaged and trusted as co-developers since the main idea of Web 2.0 is that the services and functionalities improve as the intensity of usage of these platforms rise (Bonsón, Torres, Royo, & Flores, 2012). Thus, attracting more users to be active by commenting, liking and sharing on the platform improves the social media page at hand. Furthermore, local governments need to leverage the vast social media data to improve the communication and services for the tourists of a city and to virtually manage the reputation of the city (Stieglitz & Dang-Xuan, 2013; Eshuis & Klijn, 2012).

The actions and communication of local government may greatly influence the brand perception of people, meaning local governmental city branding may affect the city's image. Local governmental city branding is defined as formal and informal communication between the local government officials and tourists about a variety of topics (Perloff, 1998/2008), such as cultural events, traffic or tourism with the aim to position a specific city on social media. An example where local governmental city branding may affect city brand image and invite tourists to the city is found in showcasing that the city has been named in a prime destination list of a trustworthy source, such as The Guardian. By doing so, the local government uses the reputation of renowned brands to boost the city's image (Braun, 2011). Consequently, it becomes likely that someone picks up this thread from the local government and shares or posts about the topic themselves. Thus, local governmental city branding may influence not only people's perception but it may also accelerate and alter the way they communicate on social media. In the remainder of this study, social media communication by tourists about a city is referred to as eWOM communication by tourists.

1.2 Goal, scope and research question

There is ample research on the role of official marketing strategies of governments (Lucarelli & Berg, 2011; Van den Berg & Braun, 1999; Braun, 2011). However, little quantitative research has been done on how local governmental city branding influences how the city brand is used by a target group in different countries. In particular, there is scarce comparative research on how local governmental city branding influences eWOM communications by large groups, such as tourists. This creates a gap in the public administration field since not all possible modes of local governmental city branding that may affect the online presence of the intended brand are considered. Therefore, the goal of this study is to add to the formation of theory on the influence of local governmental city branding on eWOM communication by tourists, by statistically testing hypothesis on this influence in Copenhagen, Denmark, and Tallinn, Estonia. Even though providing proof can be difficult in social sciences due to the lack of hard theory, compared to, for instance economics, the aim of this research is to learn from the case studies and contribute to predictive theory (Flyvbjerg, 2006).

The social media accounts analyzed are officially owned and funded by the local government of the respective city. Copenhagen is one of the most popular tourist destinations in Northern Europe (Jørgensen & Munar, 2009). Contrastingly, Tallinn, European Capital of Culture 2011,

has only recently risen to the stage as a popular Eastern European destination for tourists (Tooman & Müristaja, 2014). The popularity of both cities among tourists causes these cities to be fit for local governmental city branding analysis. This study aims to answer the following research question:

> *What is the influence of local governmental city branding on eWOM communication by tourists in Copenhagen and Tallinn?*

To answer the research question, the following sub-questions are set:

1. Which theoretical insights does the literature offer on social media communication?
2. Which theoretical insights does the literature offer on city branding?
3. How can local governmental city branding be characterized?
4. How can eWOM communication by tourists be characterized?

1.3 Relevance

There has not been any academic comparative research conducted on the influence of local governmental city branding on eWOM communications by tourists in Estonia and Denmark. The scientific relevance of this study is found in filling the gap in public administration research by connecting local governmental city branding with eWOM communications by tourists. Moreover, there is no quantitative research on how governmental city branding influences eWOM communication by tourists. It is beneficial to compare a Western European city with an Eastern European city, as such a comparison may uncover differences in local governmental city branding and social media communications cultures. Moreover, the scientific relevance of the study is to contribute to the growth of quantitative research in the social sciences field and combine the added value of statistical analysis and case study research (Flyvbjerg, 2006).

According to Lucarelli & Berg (2011) the field of city branding is highly researched but has a strong Anglo-Saxon focus. Future research avenues regarding local governmental city branding have been set to explore the social media culture of Eastern European countries and compare those with Western and Northern European countries (Bonsón et al., 2012). This study addresses these concerns and provides novel, comparative insights on city branding, as it compares a Northern European city with an Eastern European city, focusing on tourists as the target group for city branding.

This study's outcomes may assist policy makers and managers in choosing city branding strategies that positively affect the reputation of the city. Similarly, this study will allow tourists to better understand the local governmental city branding activities that they are presented with which is important to meaningfully interpret the local governmental context of city branding (Bryer & Zavattaro, 2011).

1.4 Outline

The remainder of this study is structured as follows. In chapter 2, relevant literature serving as a theoretical basis of this study is explored. The hypotheses and conceptual model are presented in chapter 2 as well. Next, the methodology of this study is explained in chapter 3. Chapter 4 covers the case descriptions, followed by chapter 5 which presents results. Chapter 6 includes a discussion of the results, limitations of the study and avenues for future research. Finally, chapter 7 presents some recommendations based on the study's results.

2 Theoretical framework

This chapter presents relevant theoretical concepts on branding, specifically local governmental city banding, and social media communication, namely eWOM communications by tourists, forming the building blocks of this study.

2.1 Branding

Governments have consciously tried to design and promote a desired image of a place since the early days of civic government, although only since around the mid-1980s branding has been generally accepted as a suitable activity for public sector organizations (Kavaratzis & Ashworth, 2005). Brands are "*symbolic constructs that add value or meaning to something in order to distinguish it from its competitors*" (Eshuis & Klijn, 2012, p. 3). In other words, brands may enhance or diminish the value of something by assigning a reputation to it. A brand cannot be fully controlled since people interpret brands according to their own mental maps and personal experiences (Kavaratzis, 2004), causing the same brand to evoke varying associations to different people. The webs of associations created in people's mind are what give meaning to a product or a place that is being branded (Klijn, Eshuis, & Braun, 2012).

Eshuis & Klijn (2012) refer to branding as a deliberate process, although public organizations, such as the tax collection service, awakening feelings of helplessness by slow processes and complicated bureaucratic procedures suggest branding may be viewed as an unconscious process as well. This way an unintended process may lead to a negative image that is difficult to change. This two-sided perspective of branding stresses the dynamic nature of brands, as well as their inherent uncontrollability. Trueman, Klemm and Giroud (2004) remark the importance of official communicated brand developed by a brand manager, such as a City Tourist Office. However, next to formal communication, there is uncontrollable tertiary communication, entailing eWOM and media communication (Kavaratzis, 2004). Steering this type of communication is complicated since there is a vast amount of intertwined communication lines between various people. Developing a city brand can involve target groups who have the capacity to either work against the whole branding process or to co-produce the brand (Klijn et al., 2012).

A brand's success ultimately relies on the users' acceptance of it, shaping both conscious and unconscious modifications of the brand's meaning, with the latter type of modification referring to the unofficial brand image that tourists interpret (Trueman et al., 2004) and promote through eWOM communication. The more the brand manager is able to create a brand through topics users relate to, the more frequently and intensely people will communicate about the brand (Klijn et al., 2012). Brand managers have it in their best interest to encourage their audience to voluntarily partake in the bottom-up co-production process (Merz, He, & Vargo, 2009). This way the people are involved in the development of the brand, effectively strengthening the brand and reducing the risk of counter branding – meaning users unveiling a negative side of the brand opposite to the intended brand image (Eshuis & Klijn, 2012).

Co-producing a brand with the support of local communities and the experience of tourists enables brand managers to reflect the desired image (Trueman et al., 2004). The destination brand is closely related to the destination image, the latter, however, is controlled by tourists (Jørgensen & Munar, 2009). The eWOM communications by tourists have an influence on how the brand is used and how the city is promoted by large groups of people. Thus, branding is not a merely a top-down process but can rather involve users (Eshuis & Klijn, 2012; Muniz Jr. & O'Guinn, 2001; Kavaratzis, 2004; Braun, 2011; Szondi, 2007; Yan, 2011). This creates a dilemma since involving tourists requires effort and coordination. However, if tourists are ignored they may reconstruct the brand image according to their experience and cause more harm with the negative exposure (Eshuis & Klijn, 2012).

2.1.1 Local governmental city branding

A city consists of various elements, such as buildings, events and people, but also pollution and traffic jams. These elements transcend logos or symbols and are what constitute the visual evidence of a city (Trueman et al., 2004) that people actually encounter. Cities are characterized by both positive and negative facets, and the ways people perceive different cities varies greatly. Cities can be branded by emphasizing their unique features. This helps to define the unique identity of cities, and to make the city stand out from its competitors (Morgan, Pritchard, & Piggott, 2002). The specific target group focus of city branding depends on the needs and characteristics of the city, implying that, for instance underdeveloped countries are more interested in gaining foreign investments, while richer countries have the opportunity to focus in-depth on tourists or residents.

Foremost, in order to gain credibility and improve the cities' experience, a brand must have depth and be built on a realistic basis by evoking strong associations connected to the concerned city (Vanolo, 2008; Szondi, 2007; Braun, 2011). A superficial or misleading brand might get one-time attention, but the reality of a city will eventually catch up with the illusion (Yan, 2011). City branding entails developing what Szondi (2007) calls a coherent marketing approach, which can be managed by various organizations, including the local government. However, the multiplicity of stakeholders is a challenge to branding (Morgan et al., 2002) since people can have different perspectives on how to reach their goals, causing uncertainty in the decision-making process. Furthermore, local governmental city branding can be rather inconsistent since local government priorities often change and the vision about a city becomes blurred (Trueman et al., 2004). This is due to regular elections, often every four years, which result in local government employees changing office frequently.

From the perspective of local government there is a need to engage with the informal communication of tourists of a city, to discuss and gain information about the public opinion to identify current issues and predict future topics (Klinger & Svensson, 2014) that can be used for city branding. Sending a great amount of minor information bites entails a risk for the local government to provide contradicting messages. Thus, communication managers need to internally inform involved people about the content or keywords and length of the posts that will be communicated to external stakeholders (Eshuis & Klijn, 2012; Szondi, 2007). This is a paramount aspect of maintaining the credibility of a public organization since it assures external parties the professional and informed style of local government. Thus, successful local governmental city branding can exploit the available communication platforms (Louw, 2005). However, it should be considered which type of post, that is video, link, status update or photo, is most suitable for the message at hand. Kwok and Yu (2013) found that statistically photos and status posts receive most likes, and status posts are the ones that have the greatest number of comments as well.

Braun (2011) highlights the development of a shared unambiguous agreement on branding goals among the governmental leadership as a key success factor in the branding process. This is important since branding can be defined in various ways and to create a strong brand the brand managers might want to work towards a similar end goal. This is even more the case with city branding, where local governments have different objectives, such as the community's wealth and well-being, as compared to businesses whose main driver is profit (Van den Berg & Braun, 1999).

Furthermore, a credible city branding umbrella vision which covers all current and potential tourists can be included in the government agenda to attract specific target groups, for instance by introducing sub-brands (Braun, 2011), such as a city card for tourists that offers discounts, for instance on events and sightseeing tours. The inclusion of particular elements attracting specific target groups places the brand in people's minds and distinguishes it from the competitiors by making the brand a prime product or symbol to the target group (Hankinson, 2001). The more branding activities are taken and the more users are engaged in the process, the more branding activities by tourists will be induced (Klijn et al., 2012). Thus, on the basis of the literature it can be expected that as local government communicates about a city on social media more frequently, more tourists will communicate about the brand on social media.

Branding can be seen as a communication tool, where brand managers, target groups and consumers engage in a two-way information flow about the brand image (Kavaratzis & Ashworth, 2005). Nevertheless, branding can also be a one-way process, for instance when people are not aware of the brand or when brands are not used. With regard to the platform where branding activities take place a similar trajectory to consumer culture is evident, where social media have become salient (Muniz Jr. & O'Guinn, 2001). Since tourists in modern societies most actively use social media platforms, brand developers might want to ensure that a brand has online visibility (Kwok & Yu, 2013).

Place branding by tourists can be especially unpredictable since the experience of tourists cannot be controlled (Hankinson, 2001). Furthermore, with the ubiquity of social media tourists can quickly share opinions and experences online giving a topic exposure and distributing it through several nodes of other people (Klinger & Svensson, 2014). Facebook has a focus on social connectedness (Smith, Fischer, & Yongjian, 2012) that invites users to get more engaged and active on the platform (Duggan, Ellison, Lampe, Lenhart, & Madden, 2015). For instance, allowing tourists to post personal information or pictures on the platform creates a sense of familiarity with other tourists (Hennig-Thurau, Gwinner, Walsh, & Gremler, 2004). The eWOM communication is a beneficial source of information for brand managers with regards to brand development (Ye, Law, Gu, & Chen, 2011). However, most online information does not have a viral effect, meaning maximum exposure, but remains unnoticed which calls for intermediaries, such as highly visited social media platforms, to act as catalysts (Klinger & Svensson, 2014).

2.2 Social media communication

Communication in modern societies is a two-layer process where first, people receive and interpret messages through diverse forms of media. Second, people engage in face-to-face communication processes in informal conversations and formal meetings by being a part of the daily life social environment (Dahlgren, 2005). Both of these communication layers are important for two reasons. First, they enable people to gather news and information on other issues from external sources, surpassing their direct environment. Second, these layers create a platform where opportunities to discuss current topics with other people arise, allowing one to sense the norms and viewpoints of others. The likelihood of communicating with a person increases by some level of familiarity, meaning that one is most likely to get in touch with friends, family or neighbours. However, local governments using social media can entail improved access and immediacy for users to directly engage in the communication process through comments (Bonsón et al., 2012), likes and shares.

As Dahlgren (2013) remarks, the essence of civic communication has its foundations in talk, far from formalized deliberation, meaning that informal communications outside official meeting rooms are highly important. Decisions are influenced by a bottom-up flow of opinions and self-expression. In modern societies, this talk in the form of online communication has become undeniably important (Chadwick, 2006; Dahlgren, 2005; Street, 2011). The development of Web 2.0, meaning user-generated content, blogging, Rich Site Summary (RSS) and so forth (O'Reilly, 2007) led the way to the development of social media (Bonsón et al., 2012). Social media is defined as Internet-platforms that enable social communication and dialogue between large groups of people (Bryer & Zavattaro, 2011).

Nowadays, diverse web platforms give tourists and organizations opportunities to horizontally connect with one another through information sharing (Dahlgren, 2013). Websites, forums, blogs and so forth allow people with similar interests to gather without the geographical constraints of the offline public sphere, adding knowledge from outside the community to the local information pool (Chadwick, 2006). Due to its fluid and open nature, the Internet becomes a potential central arena for civic communication (Dahlgren, 2005). The most popular social networking site by far is Facebook (Duggan et al., 2015) which was developed in 2004. Facebook serves as a platform for discussions and for enhancing the development of collective identities (Dahlgren, 2013). Making the personal views of people public online changes the private views into public views (Gerbner, 1969) and has the capacity to connect people with

similar point of views. Moreover, Facebook is also the most widely used platform among local governments (Oliveira & Welch, 2013).

However, online communication tends to lack meaningful dialogue. Content-rich dialogue might drive attention-poor readers away and thus is replaced by information which is packaged into easily understandable sound bites (Street, 2011). This derives from the phenomenon that electronic messages often lack the meaning and intensity of non-electronic messages (Bimber, 1999). Online communication is perceived as less official compared to, for instance sending letters by post. The purpose of social media is to offer quick and interactive information to readers with a low cost (Klinger & Svensson, 2014). As tourists use social media for informal communication, local governments adjust their regular communication to fit the medium. Moreover, the development of Web 2.0 has created opportunities for the audience to address specific niche parts of the Internet where people have ample choices of what to read online based on interest (Klinger & Svensson, 2014). This creates a scene where short and catchy information bits gain most attention. This is due to the fact that reading thorough content-rich material requires both time and effort that people might not want to dedicate to going in depth with local governmental city branding. Thus, even though social media offers the freedom of revisiting messages at different times, it is important for online information to be frequently updated (Klinger & Svensson, 2014).

Social media analysis does not focus on the information in the texts but on the larger context of individual and group messages (Gerbner, 1969). Thus, the analysis is not limited to factual statements but relies on investigating what is said, and how it is said. Sometimes merely being present online entails a simulated presence that surrenders to the closed nature of public agencies and does not entail interactivity or deliberation with tourists (Bryer & Zavattaro, 2011). For instance, there might exist a homepage for an institution but if it does not provide the reader with usable information the homepage becomes obsolete. This makes it important not only to investigate the amount of messages or topics being discussed but also to explore the content of these messages and the activity of the page. Popular topics include traffic, news, weather updates, public service announcements, food or cultural events and so forth (Kavanaugh, et al., 2012). Despite the potential to explore such topics, social media might not be used to their fullest capacity as local government-tourist relations are mainly one-way communication streams since the government does not want to give up its control over the message (Bryer & Zavattaro, 2011). From the local government perspective the two main aims

of social media analysis is to manage the reputation of a city and to monitor the user-generated content of tourists (Stieglitz & Dang-Xuan, 2013).

2.2.1 eWOM communication by tourists

eWOM communication can be defined as any opinion, regardless of the sentiment, expressed by a potential or actual tourist about a place, that is available for a large group of people through the Internet (Hennig-Thurau et al., 2004). Tourists in modern societies use the Internet as the main mode of communication (Kavanaugh, et al., 2012). Thus, if governments wish to communicate with tourists, they can also focus on the Internet as the primary medium to reach tourists through the platform that tourists are using. These platforms allow large quantities of social media data to be gathered and analyzed (Kwok & Yu, 2013). Following Gerbner (1969) message analysis can be conducted by measuring 1) the attention or topic of posts by identifying popular words, 2) the emphasis of the message, meaning the length or intensity of the text and 3) structure of the message; which words occur together frequently. This study will modify and partly apply the message analysis concept of Gerbner (1969) on Facebook analysis. Social media analysis connects the vast user-generated data of individual communication channels with local governmental city branding (Stieglitz & Dang-Xuan, 2013).

Social media allows information to be distributed in an interactive way by allowing comments, status updates, hyperlinks and adding visualizations to text with photos and videos (Dahlgren, 2013). However, this is a rather passive engagement since the shares and comments of tourists are dependent on texts and topics chosen and produced by the governmental parties (Louw, 2005). Nonetheless, social media allows users to partake in tourist-to-government communication and to influence the decisions of local government in innovative ways (Bonsón et al., 2012). Furthermore, social media invites otherwise less politically engaged citizens to express themselves online influencing the intensity of tourists' comments and the number of likes and shares on the local governmental posts (Bimber, 1999). Consequently, eWOM can be operationalized by the number of comments (Hennig-Thurau et al., 2004), likes (Kwok & Yu, 2013) and shares. Furthermore, the type of posts in Facebook can be categorized under popular clusters, such as, hyperlink, photo, video and status (Kwok & Yu, 2013), whereas events are less common and notes rarely used.

a. Type of post

Different types of posts, such as photos or videos, have a statistcally significant effect both on likes, as well as comments and shares (Cvijikj, Spiegler, & Michahelles, 2011). Research shows that posts including photos enhance the amount of likes a Facebook post receives (Malhotra, Malhotra, & See, 2013). Cvijikj et al. (2011) researched various Facebook brand pages for companies such as Coca-Cola and found that status posts, meaning posts with only text, were the most popular type. This tendency could have been due to the fact that in 2011 photos, videos and link were less common than in the following years. On the basis of the literature, the following hypothesis can be formulated and tested in the remainder of this thesis.

H1. The type of local governmental city branding influences the likes/comments/shares count of eWOM communication by tourists on Facebook.

b. Post length

Malhotra et al. (2013) find length to be an important factor influencing the amount of likes, namely it is found that posts should be as conscise as possible in order to get more likes. This fits with the general aim of social media to be instantly appealing to an audience who wants to be informed or entertained quickly. However, research is not conclusive on this matter as the length of the post has also been found to have positive impact on the number of likes (Sabate, Berbegal-Mirabent, Cañabate, & Lebherz, 2014). Thus, the following hypothesis can be formulated and tested in the remainder of this thesis.

H2. The length of local governmental city branding influences the likes/comments/shares count of eWOM communication by tourists on Facebook.

c. Year of post

A review study based on information until 2011 shows that Facebook research, measured in articles published, has annually grown in popularity from 2005 (Wilson, Gosling, & Graham, 2012). The third hypothesis will test whether the same trend applies within tourists replying to posts on Facebook. Namely, it is researched whether the year of the post has a significant influence on the likes/comments/shares of the post.

H3. The year of local governmental city branding influences the likes/comments/shares count of eWOM communication by tourists on Facebook.

d. Month of post

Another aspect that may influence tourist eWOM communication is temporality, that is the seasonal aspect of tourist behaviour (Ye et al., 2011) This makes it interesting to investigate the time of eWOM posts which can be analyzed per month, as well as day or hour of the day.

Tourism is a seasonal field and it calls for more creativity to come up with posts off-season, however brand managers are encouraged to make the effort and possibly invite people to visit a place twice (Carter, 2014). Visitors might not be aware of activities or events happening in a city off-season even though they would be interested. On the basis of the literature the fourth hypothesis can be formulated and tested in the remainder of this thesis.

H4. The month of local governmental city branding influences the likes/comments/shares count of eWOM communication by tourists on Facebook.

e. Weekday of post

De Vries, Gensler and Leeflang (2012) researched the brand posts of fans from different fields and found that the most popular day for posting is Thursday and generally new posts were made every two days. However, researchers have also found that Fridays and Wednesdays are the days where most posts are being placed (Cvijikj et al., 2011). Thus, research is inconclusive on this aspect of Facebook posts. Thus, the fifth hypothesis can be formulated and is tested in the remainder of this thesis.

H5. The weekday of local governmental city branding influences the likes/comments/shares count of eWOM communication by tourists on Facebook.

f. Hour of post

Conducting Facebook research on a daily level reveals that posting on peak activity hours of the users, between 4 PM and 4 AM, results in a higher number of likes and comments (Cvijikj & Michahelles, 2013). Thus, the sixth hypothesis can be formulated and is tested in the remainder of this thesis.

H6. The hour of the day of local governmental city branding influences the likes/comments/shares count of eWOM communication by tourists on Facebook.

2.2.2 Emotions

The final section that forms the theoretical grounds of this study deals with emotions. Following Turner and Stets (2005), emotions can be defined as socially constructed expressions of a situation through mental response or physical moves which are induced by activation of interconnected pathways of the brain. Emotions are informal social constructs in the sense that they are not officially recorded, however in various situations there are certain routine emotions that people express which are considered socially appropriate (Lowndes & Roberts, 2013). For instance, when a national tragedy happens, it is generally considered appropriate to feel either sad or compassionate. However, the emotions that are legitimatized depend on the social arena (Lowndes & Roberts, 2013), meaning that emotions are not universal.

However, emotions can be classified under what Turner and Stets (2005) call primary or secondary, as well as low, moderate or high intensity emotions. Primary emotions include for instance high intensity emotions pride and love, moderate intensity emotions friendliness, enjoyment and expectancy, as well as a low intensity emotion acceptance, or content (Turner & Stets, 2005). The lists of basic emotions can slightly vary, for instance Mohammad and Turney (2010) exclude acceptance from the basic emotions but instead identify trust as a basic emotion. Secondary emotions include, for instance, moderate intensity emotions dispirited and gratitude (Turner & Stets, 2005).

In regard to eWOM communication people have an output to express their positive or negative emotions. Generally, people prefer certain stability in their lives and when a situation or experience moves the equilibrium state of mind to either a positive or negative side a possible output is to express their emotions online (Hennig-Thurau et al., 2004). Positive emotions, such as enjoyment and loving, significantly invite a greater number of people to react to eWOM (Chan & Li, 2010). However, it is also found that both positive and negative posts can attract a great number of comments by arousing general interest (De Vries et al., 2012). On the basis of the literature, the seventh hypothesis can be formulated and tested in the remainder of this thesis.

H7. The sentiment of local government posts influences the count of likes/comments/shares of eWOM communication by tourists on Facebook.

Following Mohammad and Turney (2010), certain emotions such as enjoyment and sadness, are identified as more evocative than others, such as anticipation. However, tourists prefer to react to posts that are actually relevant for them (Sabate et al., 2014). Thus, when an upcoming event is approaching tourists who are planning to attend it might react to the corresponding posts actively. The emotion in this case might often be anticipation which contrastingly, according to Mohammad and Turney (2010), would evoke low intensity reactions. Since the research is not conclusive on this aspect, the following hypothesis is formed without predicting the direction of the influence.

H8. The intensity of emotions evoked by local government posts influences the count of likes/comments/shares of eWOM communication by tourists on Facebook.

2.3 Conceptual model

The hypotheses of this study are grounded in the theoretical framework. Figure 2.1 illustrates the conceptual model. Hypotheses do not have a plus or minus sign, as there is no expectation regarding direction of the effect; the hypotheses merely state that an effect might exist. This is due to the fact that only one predictor is numerical, length, and research is inconclusive on that variable. Categorical variables do not have a direction of the effect.

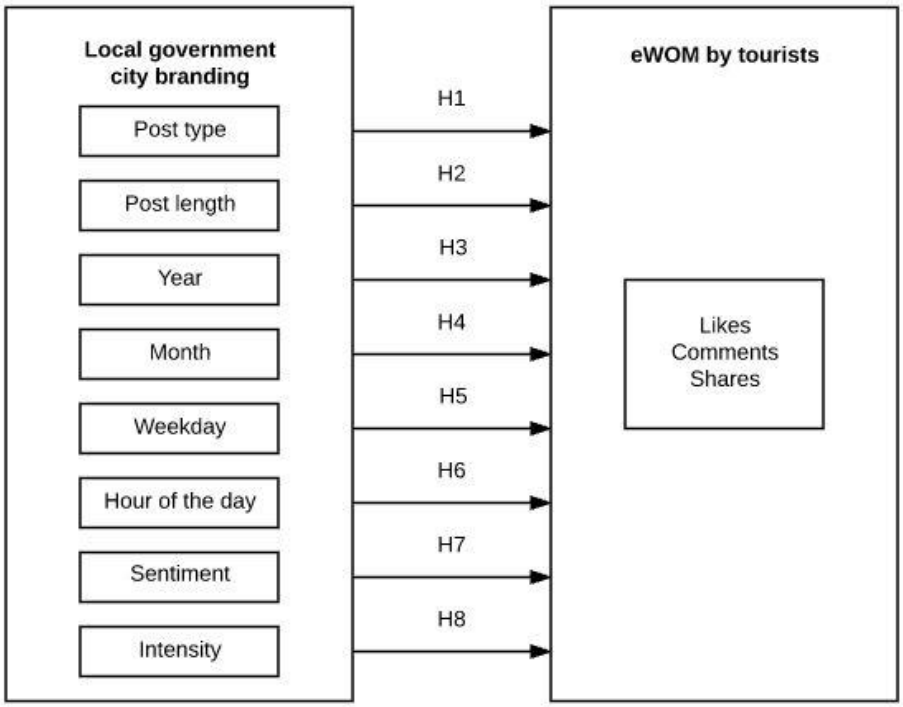

Figure 2.1 Conceptual model

3 Methodology

This chapter will shed light on the methodology of this study providing a clear and transparent overview of the steps done during the research process. First, the research design and sample of the study are explained. Second, steps of data collection, preparation and cleaning, as well as data analysis are discussed. Finally, the reliability and validity of the study are explored.

3.1 Research design

3.1.1 Large N case study

This study analyses large amounts of Facebook posts by tourists in two different cities. Thus, the research strategy is a double case study with large N of posts. The tool used to gather Facebook posts is the statistical programming language R, which has a function to scrape large quantities of raw Facebook data from a specified page. The research strategy is a case study explaining two cases, namely Copenhagen (Denmark) and Tallinn (Estonia). Researching an Eastern European and a Northern European capital city allows reducing the impact of a possible regional bias.

The selection of cases is information-oriented, as opposed to random selection, and on one hand is based on two maximum variation cases (Flyvbjerg, 2006), where Tallinn is the capital of an economically upcoming Eastern European country and Copenhagen is the capital of a Northern European fully developed state, Denmark. On the other hand, Tallinn and Copenhagen are what Flyvbjerg (2006) calls extremely good cases fit for in-depth social media analysis in the sense that both of the cities and countries have a high percentage of Internet users. Denmark is among the highest ranking European Union (EU) information society countries with around 97% Internet users from all individuals in 2015, while Estonia follows with 89% (Eurostat, 2015). The aim of choosing two cases is to avoid bias from basing the study on one country which would result in low external validity.

3.1.2 Population and sample

Quantitative Facebook text mining from two Facebook pages for a period of 81 months – from August 2009 until April 2016 – created comprehensive insights into a large set of Facebook

data with a total population of 5,299 Facebook posts. The large amount of data allows to only include posts where all the variables are present. Therefore, after excluding posts with missing values and outliers, such as one post in the total dataset with the type "note" the final sample size used was 5,168, while 4,083 posts were used for the linear regression analyses.

3.2 Research methods

3.2.1 Data collection

Data is gathered from the Facebook pages https://www.facebook.com/VisitTallinn/ and https://www.facebook.com/VisitCopenhagen/ through Facebook's Application Programming Interface (API). Exchanging one's Facebook credentials against access credentials allows one to use the API to access publicly available information (Munzert, Rubba, Meißner, & Nyhuis, 2015). Using the Facebook API to gather data is beneficial since it allows retrieving large quantities of complete data, whereas a limitation for researchers using other applications, such as Facebook Query Language (FQL), can be time limits of the past 30 days or data limits of 50 posts (Kwok & Yu, 2013).

3.2.1.1 Data collection for regression models

The government-owned social media accounts on Facebook promoting and branding the city to tourists are selected for the analysis. Facebook provided insights into 1) all local government posts pertaining to city branding and 2) tourist responses to these posts through likes, shares and comments. The analysis spanned a period from the 1 August 2009 until 30 April 2016. Appendix A1 provides an overview of the data collection output, with each post category characterized by 1) the user account it relates to, 2) the content of the original message, 3) the date and time of posting, 4) the type of post – either link, status, video, event or photo, 5) the likes count, 6) the comments count and 7) the shares count. In order to analyze in-depth how local governmental city branding posts influence the eWOM communication by tourists only likes, shares and comments to the government posts are analyzed. New posts by tourists are beyond scope of this research.

First, all Facebook posts created by the Facebook users VisitCopenhagen and VisitTallinn are retrieved (see Appendix C, line 6-38) using the `Rfacebook` package (Barbera, 2016). Mind that for each package that is not a part of the default R packages, the `install.packages()` command

with the specific name of the package has to be run once. For example, to install the `plyr` package, `install.packages("plyr")` needs to be run. Afterwards only the command `library()` with the specific package name has to be run to load the package. To replicate the code of the appendices, the `plyr` package and the `stargazer` package need to be loaded in particular, since these are used very often (see Appendix C, line 1-2). The working directory needs to be set as well, which happens on line 3 of Appendix C. Since the `Rfacebook` package is in the process of being further developed, the complementary package `devtools` (Wickham & Chang, 2016) is used to access the most recent `Rfacebook` package. 1 August 2009 has been chosen as the starting date, as both users started posting more frequently in this month. As some posts that were created in February were not captured by the initial command due to technical limitations of the Facebook API, another downloading iteration was performed which filled all the missing months (see Appendix C, line 36-38).

Downloading this data results in two separate datasets, one for Copenhagen and one for Tallinn. A city code is then assigned to posts relating to each city, with 0 corresponding to Copenhagen and 1 corresponding to Tallinn. Next, all posts are bound into a single dataset (see Appendix C, line 40-41).

3.2.1.2 Data collection for sentiment analysis

Manual sentiment analysis is done on 215 status posts after 47 posts that included an http:// link are excluded from the initial status post dataset of 262 posts (see Appendix C, line 67-69). Those posts are excluded since assessing the sentiment is hindered when an external reference of content is part of the message but out of scope of the research. Appendix A2 shows sample data with the manual analysis applied, while the code can be found in Appendix C, line 198-211.

3.2.2 Data preparation

Before data analysis there is an important process of data preparation and cleaning, which includes processes to remove unusable data, filter certain data out and streamline data (Nisbet, Elder, & Miner, 2009). The most suitable function in R for this procedure is found in the text mining `tm` package and its corresponding `tm_map()` function (Munzert et al., 2015). All posts without any message are removed, concerning 83 posts (see Appendix C, line 47). The

messages corresponding to each remaining post are cleaned in order to be able to properly analyze length of the messages (see Appendix C, line 50-58). A corpus of the messages is created using the `tm` package (Feinerer, Hornik, & Meyer, 2008). American Standard Code for Information Interchange (ASCII) characters, hyperlinks, punctuation, numbers and extra white spaces are removed, since these elements disturb the accuracy of the word count per post. This process helps reducing the noise in the data and provides a clarity of the analysis (Nisbet et al., 2009).

Several attributes are added to the dataset, all concerning independent variables used in the regression modeling. First, the length of each post is added (see Appendix C, line 59) by splitting the clean messages into separate words, using `strsplit()`. The result is turned into several strings using `unlist()`. The number of individual strings is counted using `length()`, returning the total number of words in the message. This number is added to the dataset.

Next, the hour of the day, weekday, month and year are added to the dataset (see Appendix C, line 62-65). For hour, month and year this is done using `substr()`, which retrieves a part of a string. For example, to retrieve the hour of the day, the 12th until the 13th character of `created_time` is retrieved. The variable `created_time` has the following format: `2012-01-26T11:19:00+0000`. Thus, retrieving the 12th and 13th character returns 11, which is the hour of the day. The same principle applies to month and year. This implies the minutes are simply dropped from the timestamp: for example, 14:23 and 14:58 both become 2 PM. There is a special function to retrieve the weekday name based on the date, namely `weekdays()`.

Finally, several posts are removed as they are not deemed appropriate for the analysis (see Appendix C, line 71-74). All statuses with hyperlinks in the message are removed, as they cannot be considered statuses nor links. This concerns 47 posts in total. All statuses without links in them are saved in a separate dataset since they will be used for manual analysis later on. Moreover, all posts with type `note` are removed; this is only a single post. As there is only one of this type, it is not considered necessary to keep this post.

3.2.3 Analysis

3.2.3.1 Message analysis

Message analysis can be conducted by looking at variables, such as the type of post, the number of positive and negative reactions on posts and level of activity, i.e. the comments a post

receives (Bonsón et al., 2012). In this study, the content, type, intensity, i.e. length, and timestamp of all posts created by local governments through the city promotion accounts and tourists are considered. These variables are highly relevant for the purpose of this study. Intensity is analyzed as it is a valid indicator of online activity (Bimber, 1999). The aim is to discover patterns and calculate links between data (Nisbet et al. 2009).

3.2.3.1.1 Operationalization

Table 3.1 shows the operationalization of these elements and their assigned values. The variables are measured on an individual post-level. The timestamp of the post is collected as well, where a logical structure of values is followed for hour, following a 24-hour range, and month, for instance month January was assigned value 1, while December is assigned value 12. Each post also includes a value for year, where posts from 2009 are assigned value 2009 for year and posts from other years are assigned a value of the respective year. The years range from 2009 until 2016. Weekdays are assigned values between 1 for Monday to 7 for Sunday. For clear interpretation the values of weekdays are labelled, thus weekday with the value 1 is labelled Monday, weekday with the value 2 is labelled Tuesday and so forth until the final weekday with the value 7, which is labelled Sunday. All months, weekdays and hours are included, as well as all years within the timespan of the research (2009 until 2016).

Regarding the tone, posts are labeled 0 (neutral) or 1 (positive). Since only a single post initially received the tone negative, this post is also considered as neutral. As for the intensity, posts are labeled 1 (least intense), 2 (moderately intense) or 3 (most intense).

Table 3.1 Operationalization of variables in linear regression

Concept	Definition	Indicator	Value	Level of measurement
Characterization	Categorization of posts per content	Type	Event, link, photo, status, video	nominal
Emphasis	Size of the post	Length	x words per post	count
Trend	Change over longer time period	Year	Range from 2009 – 2016	ordinal
Seasonality	Periodic change	Month	Range from 1 (January) until 12 (December)	ordinal
Activity	Daily change	Weekday	Range from 1 (Monday) to 7 (Sunday)	ordinal

Temporality	Time of the post	Hour	Range from 0 – 23	ordinal
Sentiment	Underlying emotion of the post	Tone	Neutral – 0, positive – 1	binomial
Magnitude	Strength of the emotion of the post	Intensity	1 – low, 2 – moderate, 3 – high	ordinal

By gathering the values for all elements for each post, an inference can be made about posts at an individual level. The categorization and timestamps allowed for potential trends and relationships to be discovered. Moreover, the categorization of the posts allowed for higher-level analysis, as it is possible to analyze posts individually as well as grouped, for instance looking at all the posts that are of the type "photo".

3.2.3.1.2 Selecting variables and converting data types

Next, several variables of the dataset are selected for the analysis. Some attributes, namely poster ID, poster name, message, full date and time of creation, link and post ID, will not be used for the analysis. Thus, a new dataset is created where these attributes are omitted. This is done by selecting only the desired columns (see Appendix C, line 77-78).

Various attributes are changed from data type character to data type factor, so that R can recognize the attributes as categorical variables in the regression modelling. This is done for type (column 1), city (column 5), hour of the day, weekday, month and year (columns 7 through 10). The dataset in this form is copied in order to create plots with it later on (see Appendix C, line 79-80). All the categorical variables were coded into dummy variables, meaning a baseline group of the variable is assigned the value 0 and other groups are coded with a 1 (Field, 2013). The dummy variable coding is done automatically in R when using the `lm` regression function.

3.2.3.1.3 Cramer's V correlation coefficient to test for multicollinearity in predictors

Since all correlation testing is conducted on variables with more than two categories, Cramer's V is used to measure the relationships between variables in order to check for multicollinearity (Field, 2013). The contingency table is explained in section 5.2.2.7.

When nominal predictors have more than two categories Cramer's V is a suitable statistic for an effect size, or "*the strength of a relationship between variables*", which measures the

correlation coefficient, r (Field, 2013, p. 79). The contingency table values with r = .10 show a weak relationship between variables and r = .30 show a moderate effect size between variables (Cohen, 1992).

The R package `vcd` computes a contingency table of association statistics, including Cramer's V, with the command `assocstats()` (Meyer, Zeileis, & Hornik, 2015). A matrix with 7 rows and 7 columns is created with the `matrix()` command since there are 7 predictors. The results are rounded to two decimals with the `round()` command. The `stargazer` package (Hlavac, 2015) is used to create an HTML file of the output that is shown in the results. See Appendix C, line 94-97 for the complete Cramer's V code that was used.

3.2.3.1.4 Principal component analysis

Principal component analysis (PCA) can be used to transform data into linear components in order to identify whether an underlying target exists, which allows reducing the size of the data while including a great proportion of the original information (Field, 2013). The Kaiser-Meyer-Olkin (KMO) test was run with the `KMO()` command from the R package `psych` (Revelle, 2016) in order to evaluate whether the variables are suitable for running a PCA test (see Appendix C, line 83-90). If that is the case, then the scores resulting from the PCA – a weighted combination of likes, comments and shares – are used as the target variable in the linear regression models. Further assumptions of the PCA test are explained in section 5.2.1.3.

3.2.3.1.5 Regression analysis

The principal aim of linear regression is to identify predictors (X_n), where a change in the predictors, or independent variables, has an impact on the target variable (Y) (Nisbet et al., 2009). A unit change, interpreted through a measurement level, in the predictor represents the resulting change in the target (Field, 2013), which can also be seen as a change in percentage points. The data in this study has a causal link between the target and predictors as likes, comments of shares of a post depend directly on the content or structure of the post. With liking, sharing or commenting on a post tourists express their opinion about the specific post or topic. Thus, multiple linear regression is a suitable model for this study. As a result of the PCA, the three dependent variables (likes, comments and shares) are combined into a single target and used for the regression analyses.

The following equation describes multiple regression with n predictors, where b_0 stands for the constant, or intercept, of the model, b_n stands for the coefficient of predictor X_n and ϵ stands for the error term (Field, 2013, p. 298):

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \ldots + b_nX_{ni} + \epsilon_i$$

To research the magnitude and direction of the impact of each individual predictor, a simple linear regression analysis is run for each predictor. This method is applied due to the fact that the regression intercept is the mean of the baseline category in the regression which is straightforward in a simple linear regression (Field, 2013). However, in a multiple linear regression it is difficult to interpret the intercept that is combined of all the underlying baselines. Thus, simple regression is used as a supplementary technique to the multiple linear regression, since interpreting merely the results of a multiple regression model might not reveal the specific impact of different categorical variables in the data (Lord, 1967). Simple linear regression models are also applied to selected elements of the 215 status posts, namely tone and intensity.

Separate multiple linear regressions are done for Copenhagen and Tallinn separately as well, in order to identify commonalities and differences. The target variable used in the linear regression models is scaled to a range of 0 until 1. Since the multiple regression model is slightly tailed it is suitable to use normalization on the target variable in order to accurately interpret the intercept (Field, 2013).

3.2.3.2 Manual sentiment analysis

3.2.3.2.1 Coding

First, posts are assigned a topic based on phrases or words from the original message, instead of the clean message output that was used in the automated analysis above. A data-driven approach is used where topics emerge from the posts (Saldaña, 2013). The phrases which determined the decision of being assigned a topic are recorded in a separate column in the data analysis document. All messages are given one leading value for a topic, even though sometimes a message includes several topics. For instance, when a post mentions good weather as well as recommending a horse racing event for an outdoor activity, the post is categorized under the code event instead of weather as the message of the post is about the specific event.

Second, after the initial analysis of the text by the first reader the topics are examined once more and clustered under broader codes. This method is what Saldaña (2013) calls subcoding, which allows enriching the analysis by adding supplementary details to the data and is especially suitable for content analysis. After going through all the posts the coder had a more comprehensive overview of the analysis and some topics are merged under an overarching code. For instance, two posts that initially receive a topic "season" based on mentioning "summer" or "autumn" are later clustered under "weather". Descriptive statistical analysis is run on the codes, including frequency counts, which is a recommended method while using sub-coding (Saldaña, 2013). The topics are not used in further analysis as the codes are the overarching values.

Third, magnitude coding is used to include additional sub-codes to show the intensity of the sentiment of the post (Saldaña, 2013). The intensity is measured on a 3-level scale of negative, neutral and positive. All posts receive a value assigned, meaning no posts are assigned "missing". In addition, the usage of emoticons and exclamation marks, as well as the usage of caps lock is separately specified and taken into consideration when assigning the intensity values for the posts. These steps allow getting a comprehensive overview of the posts and reveal possible underlying tones, such as sarcasm.

3.2.3.2.2 Sentiment analysis

Further manual sentiment analysis is conducted on the 215 status posts to reveal the specific underlying emotions of the posts. Data-driven emotion categorization is conducted combining two emotion lists which results in eight emotions. The list includes low intensity emotions expressing satisfaction (1) "content" or showing disappointment (2) "dispirited", average intensity emotions, including (3) "friendly", (4) "enjoyment" and (5) "gratitude", as well as high intensity emotions (6) "pride" and (7) "loving" (Turner & Stets, 2005). The final emotion (8) "anticipation" was added from another basic emotion list (Mohammad & Turney, 2010) to complement what Turner and Stets (2005) call "expectancy". Exploring two sets of emotions made the choice of sentiments reliable since the majority of basic emotions were identical.

In order to ensure reliability of the analysis a second coder went through the assigned values of the posts and commented where disagreement arose. A 90% threshold is considered as a valid inter-coder agreement percentage (Saldaña, 2013). The agreement between coders in this study is 95%, resulting from agreement over 204 posts, making the coding reliable. The 5% of

the posts where disagreement arose are collaboratively re-evaluated and codes are adjusted after an intensive discussion.

3.3 Validity and reliability

Scientific research can be assessed based on the reliability and validity of the study. Reliability characterizes the ability of the measure and methods to produce identical results under the same conditions at a different time (Field, 2013). The usage of publicly accessible Facebook data and appliance of reproducible R coding methods from the beginning of the data analysis allow future researchers to replicate this study. Furthermore, as there are various ways of using R, as well as applying different statistical models, providing the full R code in the study ensures a straightforward understanding of the research process. Elaborate description of methods contributes to transparency and reliability of this study. Moreover, the large Facebook data population contributes to the study's reliability.

Internal validity stands for the degree to which a variable measures what it was actually intended to measure (Field, 2013). Table 3.1 shows the operationalization of variables exemplifying the measurability of the variables. However, some mediating or moderating variables may have been overlooked as these variables are not grounded in theory or are unmeasurable. The quantitative analysis of the posts caused limitations in how concepts, such as trends or perceptions could be operationalized. Thus, the quantitative analysis allowed for high reliability but implied limitations in internal validity.

In regard to external validity, generally whereas the benefit of case studies lies in-depth compared to for instance surveys, case studies lack breadth (Flyvbjerg, 2006). The external validity of this study is low in the sense that only two cities are included in the analysis, although regional bias is avoided. An identical method can be applied to other cities, however factors such as Internet access and cultural context should be taken into consideration (Hennig-Thurau et al., 2004). All in all, the study's validity and reliability properties are been addressed.

4 Case description

4.1 Tallinn

Tallinn, with a population over 400,000 inhabitants, is the capital of Estonia which is geographically the most northern Baltic state with a population of 1.3 million (Statistics Estonia, 2015). With close to a third of the Estonian population living in Tallinn, the city is regarded as a center for jobs, cultural events and tourism. The Republic of Estonia was declared on 24 February in 1918 but was occupied by foreign rule, mainly the Soviet Union, during and after WWII. Estonia regained independence on 20 August 1991 (Tooman & Müristaja, 2014). Estonia became an EU member state in 2007 and joined the euro zone in 2011 (Statistics Estonia, 2015).

In 2014, Estonia received over six million foreign visitors from whom more than half stayed only for one day (Statistics Estonia, 2015). This is explained by the central location of Estonia as it lies between Russia, Latvia and by boat Finland. Especially Finnish tourists can easily visit Tallinn for a day due to the regular boat traffic between Finland's capital Helsinki, and Tallinn. These neighboring countries make up two-thirds of all foreign visitors staying for at least one night in Estonia, whereas in 2014 the share of Finnish tourists was 46% (Statistics Estonia, 2015). This can be explained by the fact that Estonia has remarkable differences in population density, tourism attractions, entertainment opportunities and so forth between the capital and a few other bigger cities, such as the student capital Tartu and summer capital Pärnu, and the rest of the country. Visiting from Finland is one of the most convenient ways to reach Tallinn, compared to for instance visiting from Latvia by land, which would imply driving around 200 km from the southern border through the country up north.

The Tourist Office and Convention Bureau, which operates within the Enterprise Department, is managed by Tallinn City Government (Tooman & Müristaja, 2014). The Tourist Office and Convention Bureau aims to promote Tallinn as a tourist destination for domestic and foreign visitors in order to increase tourism revenue and employment rate (VisitTallinn, 2016). Furhtermore, the organization runs an interactive webpage www.visittallinn.ee which offers viewers recommendations on sights, restaurants, activities and areas in Tallinn worth visiting. VisitTallinn is present on several social media platforms, including Twitter, Facebook, YouTube, Flickr and Instagram. Furthermore, the Tourist Office and Convention Bureau of

Tallinn offers a freely accessible Media Bank (http://mediabank.visittallinn.ee/) where one can download and use images and videos of Tallinn with the aim to promote Tallinn as a destination.

4.2 Copenhagen

Copenhagen, with a population over 750,000 inhabitants, is the capital of the Scandinavian country Denmark with a population of slightly under 5.7 million (Statistics Denmark, 2015). Thus, Copenhagen is home for slightly over 13% of the population. The population density of 132 inhabitants per km^2 is low compared to the Netherlands with 497 inhabitants per km^2 (Statistics Denmark, 2016), but high compared to Estonia with 30 inhabitants per km^2 (Statistics Estonia, 2015). Much like Estonia, Denmark has only one metropolis, Copenhagen, which makes it an important focus for local governmental city branding (Jørgensen & Munar, 2009). In 2015, 49 million tourist overnight stays were spent in Denmark, while the main foreign tourists are Germans (Statistics Denmark, 2016).

The interactive webpage VisitCopenhagen.com, including their Facebook page, is run by the official Copenhagen Convention Bureau Wonderful Copenhagen (WoCo) (VisitCopenhagen, 2016). WoCo contributes to the development of the local tourism, for instance through cooperation with public and private stakeholders, in order to improve quality of life and the image of Copenhagen (Jørgensen & Munar, 2009). VisitCopenhagen's Facebook page with the count of 23rd of May 2016 has 113,191 likes, while VisitTallinn's page has 29,616 likes.

Copenhagen is one of the most often mentioned cities of Denmark which in 1989 led to the development of the private-public partnership Copenhagen Convention Bureau Wonderful Copenhagen (WoCo), aiming to improve Copenhagen's attractiveness primarily for tourists (Jørgensen & Munar, 2009). WoCo runs the official Copenhagen website, http://www.visitcopenhagen.dk/, which provides information about major events and festivals accessible for tourists, sights, places to eat, shopping and so forth (VisitCopenhagen, 2016).

Similarly to VisitTallinn, VisitCopenhagen is present on several social media accounts. Moreover, the Copenhagen Media Center (http://www.copenhagenmediacenter.com/) offers free photos, videos and information about Copenhagen.

5 Analysis

5.1 Descriptive statistics

After the initial data preparation and cleaning, 5,168 observations remain. Some descriptive statistics of the data are displayed below, starting with an overview of the target variables – likes, comments and shares – and post length. The summary statistics are generated and converted to HTML format using the `stargazer` package (Hlavac, 2015), while figures are created using the `ggplot2` package (Wickham & Chang, 2016). The complete R code for the descriptive statistics can be found in Appendix C, line 134-184.

Comparing the standard deviation, meaning square root of the variance, and the mean, or central tendency, of the data reveals whether the data points are widely spread from the mean (Field, 2013). The targets, including likes, comments and shares, have a relatively high standard deviation compared to the mean, whereas the predictor variables length has a relatively small standard deviation compared to the mean. Thus, the target variables are widely spread around the mean, meaning data points lie further away from the mean and the predictor length data points lie closer to the mean (Field, 2013).

Table 5.1 Descriptive statistics

Variable	Mean	St. dev.	Min	Max
Likes	246.1	966.3	0	45,895
Comments	8.6	23.8	0	836
Shares	32.8	83.1	0	2,995
Length	32.6	20.2	0	296

Figure 5.1 displays the frequencies of likes, comments and shares in absolute values per post. The quantity surpasses 500, but the x-axis of the figure has been reduced so that the most frequent quantities are properly visible. The figure shows that likes, comments and shares most frequently occur up to 50 times, after which a steep decrease takes place.

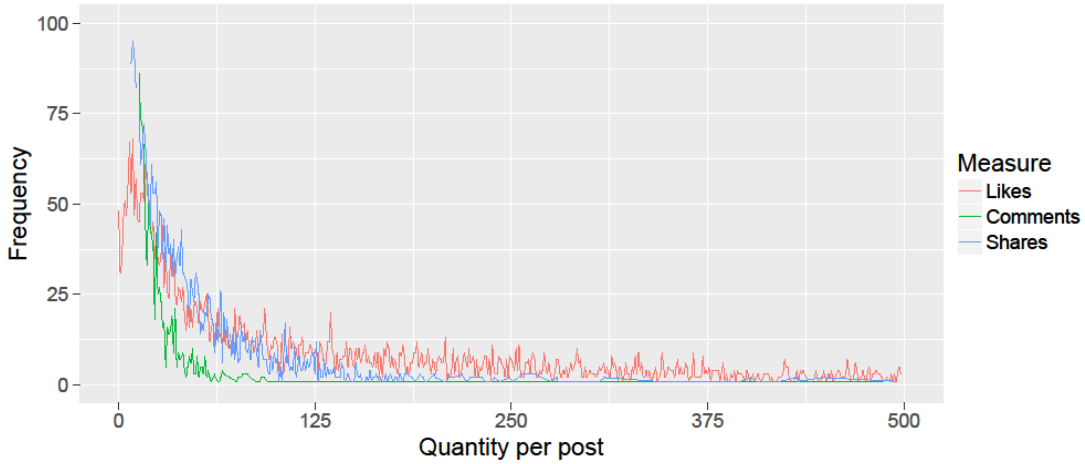

Figure 5.1 Frequencies of likes, comments and shares

Figure 5.2, the distribution graph of the predictor length shows a positively skewed data which tails off to the right side. This implies that there might be a problem with outliers, which will be discussed later in the study. The most frequent posts lengths are 16 (140 times), 21 (135 times) and 23 (132 times). Again, the x-axis has been reduced for visibility reasons.

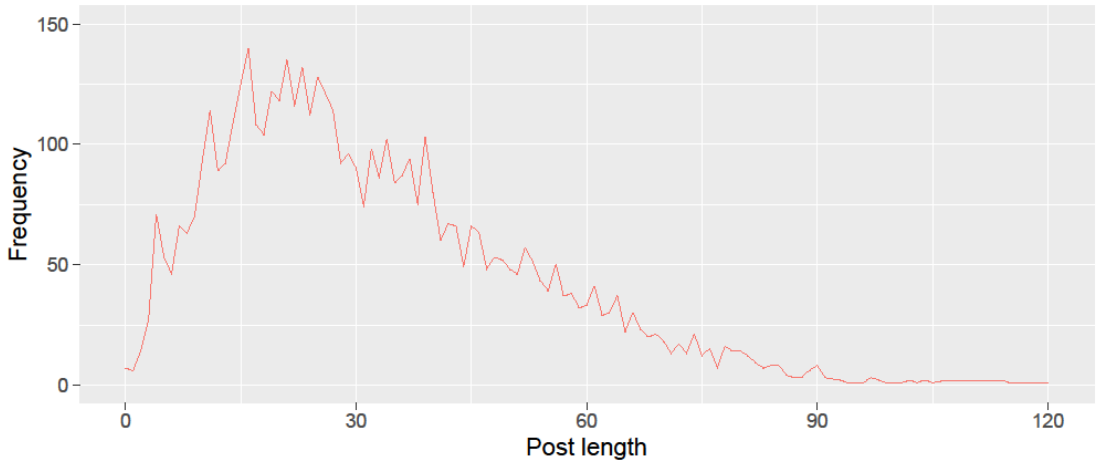

Figure 5.2 Frequencies of post lengths

For city, hour, weekday, month and year, the frequencies per value are counted using `count()` of the `plyr` package (Wickham, 2016), while the figures are created using the `ggplot2` package (Wickham & Chang, 2016). A total of 3,097 posts (60% of the total) are created by VisitCopenhagen, while 2,071 posts (40%) are created by VisitTallinn. In order to evaluate the frequency distributions, histograms are shown below (Field, 2013).

Figure 5.3 shows that the most popular hours for posting are between 2 PM and 6 PM, varying from 400 to 419 posts, with 5 PM until 6 PM being the most popular time of posting. After this time period the posting frequency starts dropping with reaching a low point at 2 AM, when only 5 posts were published. In the morning the most popular time of posting is 8 AM when in total 382 posts were published.

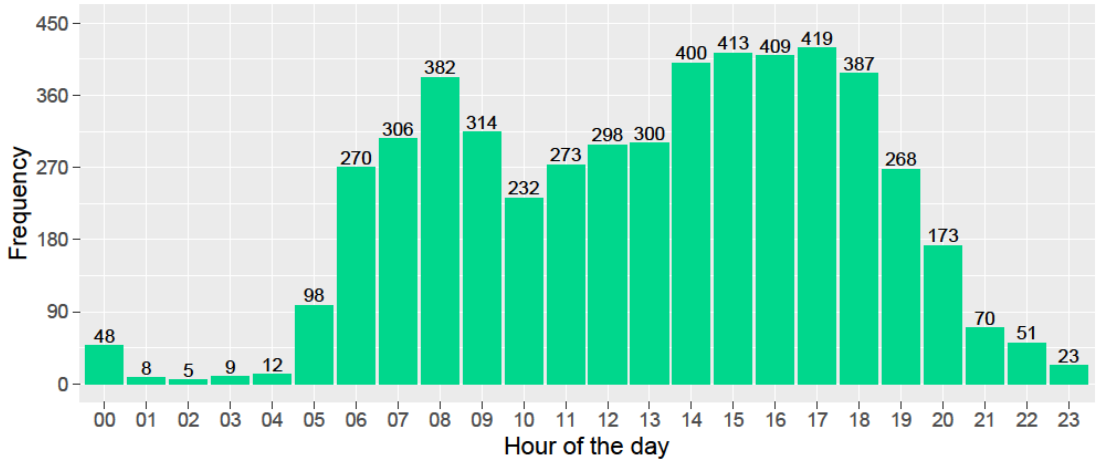

Figure 5.3 Distribution of post creation time per hour

In regard to the weekday the most posts are published on Mondays, 908 posts in total, and the least active posting days are Wednesday and Thursday, with 574 and 567 posts.

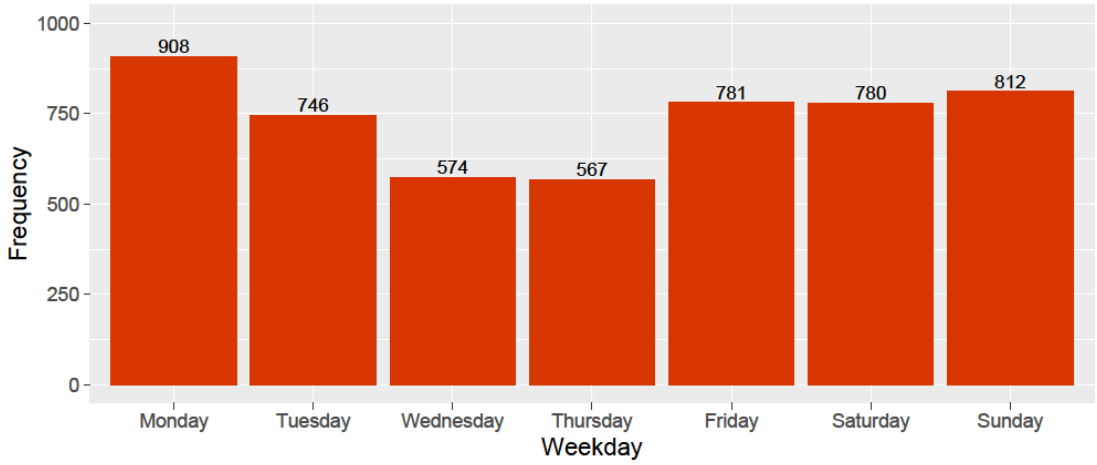

Figure 5.4 Distribution of post creation time per weekday

Examining the frequency plot of the month reveals that January and March are the most active months for posting, with 507 and 503 posts respectively. The lowest quantity of posts are published in the summer months, ranging from a minimum of 338 post in June to 395 and 390

posts in May and July, respectively. The majority of the year post distribution between months is relatively even, ranging from 413 in February to 455 in April.

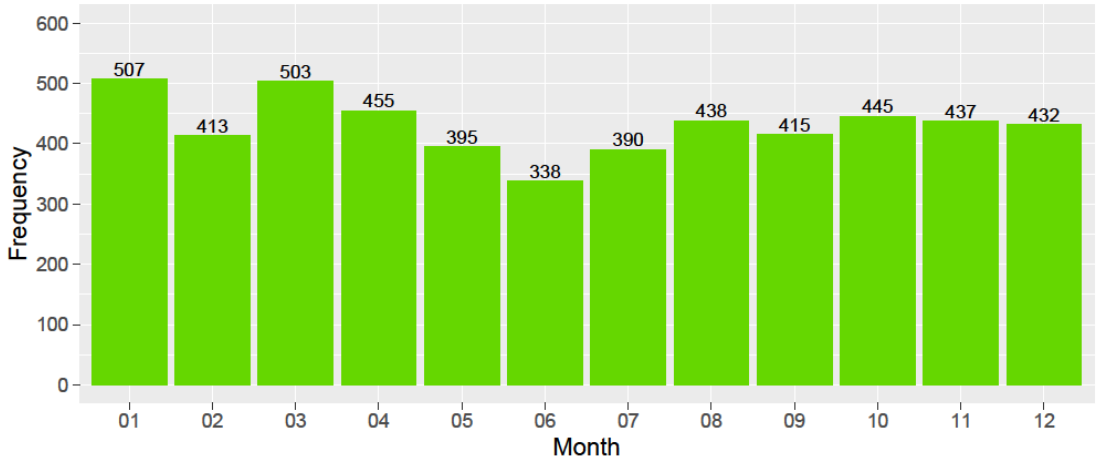

Figure 5.5 Distribution of post creation time per month

Examining the data on a yearly basis shows a trend of growing amount of posts from 2009, 136 posts, with reaching a peak in 2012 with a total of 1045 posts. Starting from 2012 the trend moves down again until 2015, where 892 posts are published. However, it must be noted that for 2009, as well as 2016 only months between January and April were included which influences the distribution. In case 2016 follows the current trend, it may be the most popular year of posting so far.

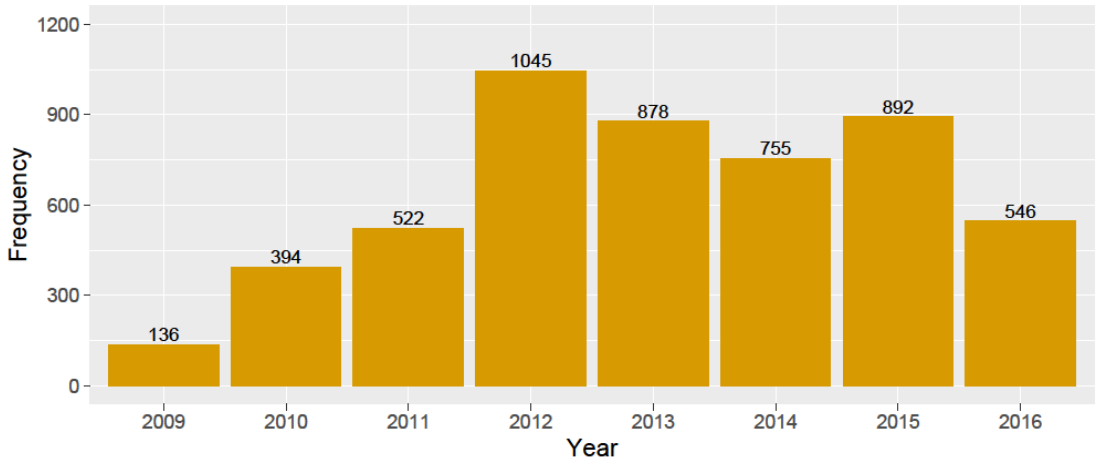

Figure 5.6 Distribution of post creation time per year

Figure 5.7 shows that by far the most popular type of post is photo with 2989 occurrences in total, followed by link with 1536 posts. The three other types, video, status and event, are marginal, with posts ranging from 117 to 286, compared to the first two categories.

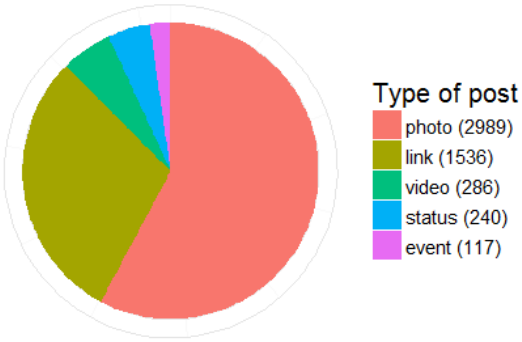

Figure 5.7 Distribution of posts per type

33 posts are considered neutral and 182 posts are considered positive. 35 posts have an intensity score of 1, 102 posts have an intensity score of 2 and 78 posts have an intensity score of 3. The mode of intensity of posts is 2 representing moderate strength of underlying emotions. More on the distributions of sentiment codes can be found in section 5.4.

5.2 Explanatory analysis

5.2.1 Principal Component Analysis

5.2.1.1 Kaiser-Meyer-Olkin test

The Kaiser-Meyer-Olkin (KMO) sample size test shows an overall measure of sampling adequacy with an acceptable threshold of .5 verifying that the variables are suitable to continue with the PCA analysis (Field, 2013). The KMO test for the main data has a result .75, which is well above the minimum acceptable threshold, therefore PCA is conducted.

The KMO test for the sentiment analysis data has a result of .56 showing that likes, comments and shares in the dataset of 215 status posts may be suitable for conducting PCA. However, the results is not significantly higher from the minimum acceptable threshold .5 implying a potential problem.

5.2.1.2 Pearson's correlation table of numeric targets

In order to increase the reliability of this study the different dependent variables, or targets, can be combined into one latent dependent variable which measures the dimensions underlying the current targets. This can be done with PCA which resulted in a sufficiently strong outcome for the main data, namely one underlying dimension was revealed which explains 87% of the variance. For the sentiment analysis data the correlations between target variables were not high enough to reveal an underlying dimension. Further, the detailed process of choosing PCA is explained.

An important element that led to choosing PCA over a similar technique factor analysis is the fact that extreme multicollinearity ($r > .8$) poses a problem for factor analysis but not for PCA (Field, 2013). All three targets in the PCA showed $r = .8$ which is considered within the upper limit of measuring multicollinearity. Furthermore, the determinant of the correlation matrix is .0918 which is much greater than the threshold of .00001. However, three variables having a correlation of more than .7 causes a multiple correlation which can be more harmful compared to a bivariate extreme correlation of .9 (Rockwell, 1975). Thus, due to multicollinearity PCA is chosen as the variable reducing technique. Bartlett's test of sphericity is significant with a *p*-value of .000, showing that variables correlate with one another significantly different from zero (Field, 2013).

Table 5.2 Correlations of dependent variables in main data set

	Likes	**Comments**	**Shares**
Likes	1		
Comments	0.80	1	
Shares	0.85	0.77	1

In regard to the data of the sentiment analysis the target variables are not highly correlated. The highest correlation is found between likes and shares, r = .36, however this is what Field (2013) calls a low communality. Thus, for the sentiment analysis data PCA is not used.

Table 5.3 Correlations of dependent variables in sentiment analysis data

	Likes	Comments	Shares
Likes	1		
Comments	0.20	1	
Shares	0.36	0.15	1

5.2.1.3 Component extraction: eigenvalues and component score

The following chapter explains the process of conducting PCA analysis. PCA is run on likes, comments and shares count (see Appendix C, line 87-89) resulting in a combined principal component with an eigenvalue 2.6. According to Kaiser (1960) a principal component with an eigenvalue over 1 results in a positive data reliability test. The eigenvalue is retrieved by taking the square root of the standard deviation of the principal component (van den Boogaart, Tolosana, & Bren, 2013). The variance proportion of the principal component is 87%, meaning 87% of the data is explained by the principal component. Since the aim of the PCA is to reduce the large set of targets and the extracted component includes the component score, all further analysis can be conducted using the component score (Field, 2013). The scores resulting from the PCA are a weighted combination of likes, comments and shares, and these scores will be used as the target variable in the linear regression models. The target variable is then normalized so that the lowest value found is 0 and the highest value found is 1. The purpose of normalizing is to make the results of the regression models easier to interpret. A sample of the resulting data as it is used for the linear regression can be found in Appendix A3.

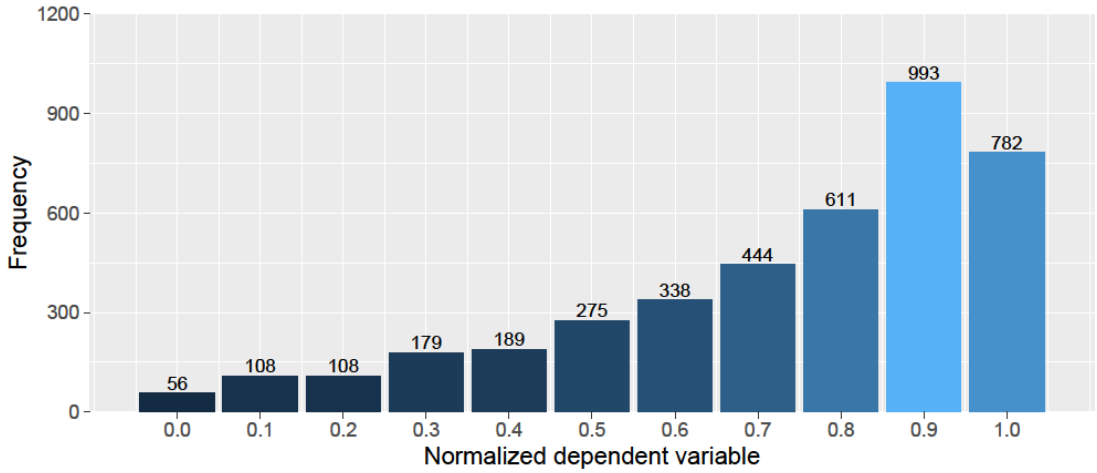

Figure 5.8 Frequencies of value normalized dependent variable

Figure 5.8 shows the distribution of the dependent target variable. The plot shows that there are many posts with a variety of low target values and many posts with a high target value. See Appendix C, line 188-195 for the code creating the plot.

5.2.2 Conditions for multiple linear regression analysis

The necessary tests were run and the assumptions for conducting multiple linear regression analysis were not violated

5.2.2.1 Theoretically informed choices of variables

The first assumption to be considered prior running a regression analysis is that the variables for regression models should be based on theoretical foundations that have been tested by researchers in the past (Field, 2013). The variables chosen for the current study are theoretically informed and explained in section 2.2.1. It is important to make theoretically informed choices before conducting a regression analysis since models are able to produce results even from variables that might not make sense in reality. This assumption is met for this study.

5.2.2.2 Measurement level

The predictor variables in a multiple regression must be numeric or binary, and the target should be quantitative (Field, 2013). The predictor length is a quantitative count variable, all other predictors are categorical variables coded into dummy variables which have two categories, 0 and 1. The target variable is a quantitative count variable. Thus, this assumption for multiple regression is met.

5.2.2.3 Normal distribution of residuals

The next assumption concerns a normal distribution of residuals (see Appendix C, line 118-129 for code regarding this assumption and other assumptions). To start with, the residuals are standardized, meaning the residuals were centered to have the standard deviation 1 and mean 0, resulting in z-scores (Field, 2013). This process does not change the shape of the histogram but provides better readability of the frequency distribution. Figure 5.9 reveals a slight negative skew of the data where the tail points to the lower scores (Field, 2013).

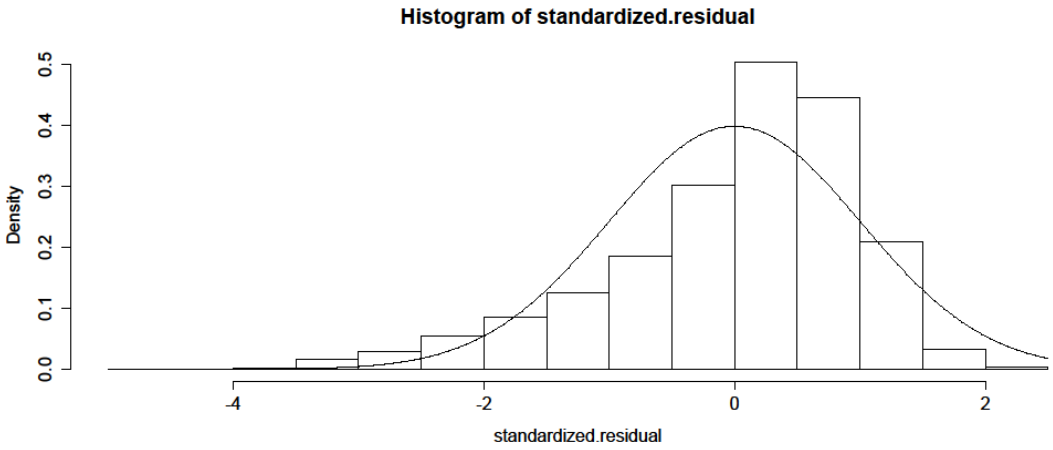

Figure 5.9 Frequency distribution of the standardized residuals

A quantile-quantile plot, or Q-Q plot, is used to further check the distribution of residuals. As Q-Q plots are based on quantile values instead of including each data point, it is a suitable normal distribution measure for large sample sizes due to easier interpretation (Field, 2013). Figure 5.10 shows that the tails of the residual lie below the line, especially on the left side of the scale. This shows a positive kurtosis, meaning the data is slightly pointy and has a disproportional amount of scores in the tails (Field, 2013).

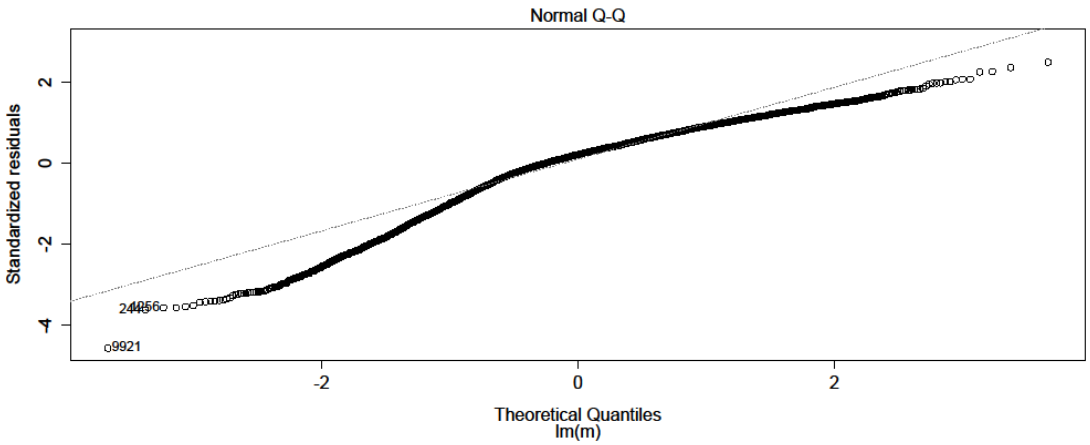

Figure 5.10 Q-Q plot of standardized residuals

However, as Field (2013) specifies, it is important to note that for large N (greater than 30) studies, according to the central limit theorem: despite a skewed shape of the population the parameter estimates will be normally distributed and the normality assumption is met. In the current study there are 136 times more samples than the acceptable threshold of a large sample

size. Moreover, these include only complete observations, meaning the posts with missing values are already excluded, which is strongly supporting the assumption of central limit theorem.

5.2.2.4 Independent errors

Residuals should have serial independence in order to ensure that the significance tests of the model are not corrupted (Field, 2013). This can be tested with the Durbin-Watson test, which ranges from 0 to 4, where a value of 2 allows to confirm that the residuals are not auto-correlated (Durbin & Watson, 1951). The Durbin-Watson test for the model shows a rounded value of 2 and a high p-value of 0.19, meaning that there is not an auto-correlated pattern in the residuals.

5.2.2.5 Outliers

Even with the application of the central limit theorem, the linear regression model is sensitive to extreme cases. To prevent extreme cases from corrupting the model a percentage based – for instance 5% or 20% – data trimming can be conducted which removes an equal percentage of outliers on both sides of the data (Field, 2013). A threshold of 10.5% proved to be the most suitable cutoff point for the data in this study as it ensures a normal distribution of the residuals in the regression model. The first and third quartile, -0.10744 and 0.14952 respectively, are in the range of ±1.5 compared to the residual standard error (yhat, 2013), 0.2169, of the model. As a result, 4,083 posts out of 5,168 remain. The code for removing outliers can be found in Appendix C, line 103-104.

5.2.2.6 Influential cases

In addition to checking for outliers in the model, it is possible to test for influential cases in the model that would alter the results significantly in case those occurrences are excluded from the model (Field, 2013). The Residuals vs. Leverage plot identifies influential cases which would be in the lower right or upper right corner outside Cook's distance; patterns are not a point of interest in this plot (Kim, 2015). Figure 5.11 shows that Cook's distance line is barely visible and there are no data points in the upper right or bottom right corner implying that there are no

influential cases in the model that would significantly affect the results. Furthermore, as a Cook's distance value over 1 implies a potential problem (Field, 2013), the Cook's distance value of -4 on Figure 5.11 confirms that there is not a problem with influential cases. Therefore, no additional cases are removed from the analysis.

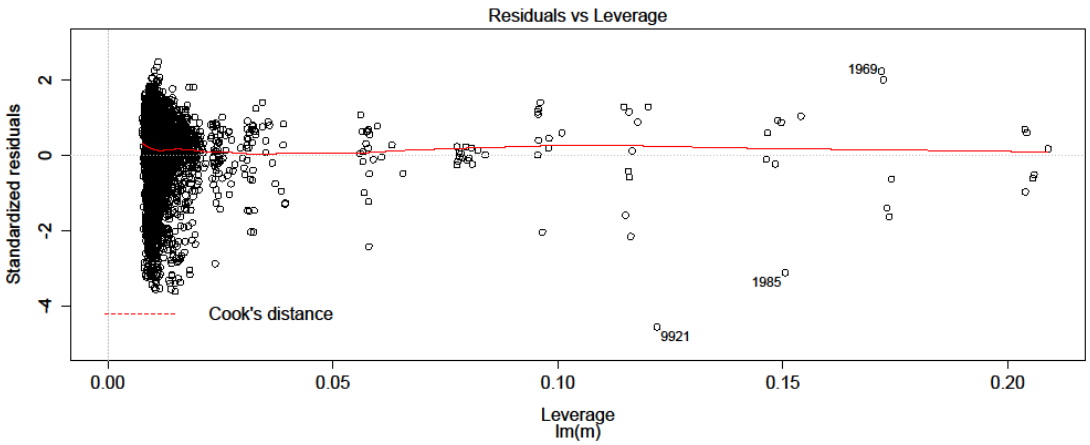

Figure 5.11 Residuals versus leverage

5.2.2.7 Assumption of no multicollinearity

The Cramer's V correlation coefficient (see Appendix C, line 94-97) results show that the predictors are measuring different concepts with the majority of $r \leq .23$. The highest effect size is evident between predictors year and type with $r = .31$ showing a moderate relationship between the variables. These results allow all predictors to be used individually in the regression models.

Table 5.4 Cramer's V correlation coefficients of independent variables

	Type	**City**	**Length**	**Month**	**Hour**	**Year**	**Weekday**
Type	1						
City	0.20	1					
Length	0.18	0.20	1				
Month	0.07	0.06	0.17	1			
Hour	0.14	0.20	0.19	0.14	1		
Year	0.31	0.25	0.23	0.20	0.21	1	
Weekday	0.12	0.04	0.16	0.06	0.13	0.06	1

Furthermore, the variance inflation factor (VIF; see Appendix C, line 127-129), which indicates potential problems of collinearity, and tolerance statistic, which is calculated by dividing 1 with VIF (Field, 2013), are explored. Specifically, generalized variance inflation factor (GVIF) is used which is a more suitable measure for models including dummy variables (Fox & Monette, 1992). The tolerance calculations (1 / GVIF) are added as a separate column. A greatest VIF value of over 10 or the mean of VIF values over 1 are indications of serious problems, while tolerance under 0.2 is a cause for concern (Field, 2013). As the GVIF values are a modification for the VIF, the general guidelines of the latter will be applied in this study.

The largest GVIF is 3.2 which is well below 10 and does not show a problem with collinearity in the model. All tolerance statistics are above 0.2 showing no sign of multicollinearity. The average GVIF is 1.7 which is not much higher than 1. Therefore, in combination with the Cramer's V correlation coefficient results above the no multicollinearity assumption is met.

Table 5.5 GVIF and tolerance statistic

	GVIF	Df	Tolerance
Type	1.636	4	0.611
City	1.185	1	0.844
Length	1.245	1	0.803
Month	1.655	11	0.604
Hour	1.933	23	0.517
Year	3.192	7	0.313
Weekday	1.127	6	0.887

5.2.2.8 Assumption of linearity and homoscedasticity

The standardized residuals vs. standardized fitted predictor, or zresid (y-axis) vs. zpred (x-axis) scatterplot shows that the assumptions of linearity and homoscedasticity are met as the scatterplot does not have a strong curve nor a funnel shape (Field, 2013). Figure 5.12 shows a slight underlying funnel shape, however there are plentiful of value outside the shape that even out the overall plot. The red line on the graph shows how much the model's prediction deviates from zero, showing the inaccuracy of the prediction. Figure 5.12 shows that the red line in

general lies close to the zero line showing an acceptable level of accuracy. While the data points on the positive y-axis become more concentrated as moving further to the right on the x-axis, this trend is compensated on the negative y-axis, where data points are widely scattered around the right side of the x-axis. The data points on the y-axis on the left side of the x-axis are quite symmetrical. Therefore, showing that generally the graph is relatively symmetrically distributed.

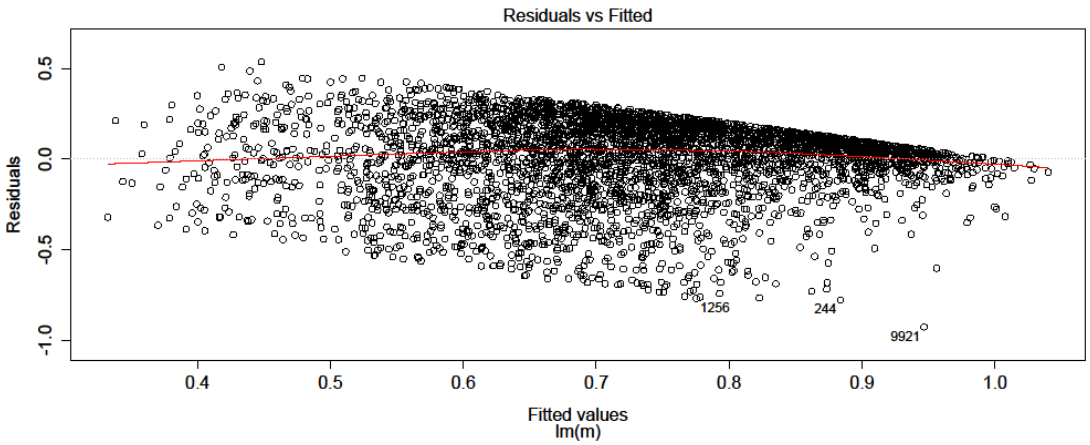

Figure 5.12 Zresid vs. zpred

5.2.3 Multiple regression model

5.2.3.1 Model summary

Since all the assumption for multiple regression are met, the model is run (see Appendix C, line 101-112). The variables are entered into the model through forced entry, meaning all variables are entered at the same time, which is considered an appropriate technique for theory testing (Field, 2013). As stated before, the target variable is scaled so that the lowest value found is 0 and the highest value found is 1. Further, the accuracy of the model is examined. Table 5.6 shows the summary results of the multiple linear regression.

Table 5.6 Multiple linear regression summary

Residual standard error: 0.22 on 4029 degrees of freedom	
Multiple R-squared: 0.26	Adjusted R-squared: 0.25
F-statistic: 27.29 on 53 and 4029 DF	**p-value**: < .001

The summary fraction of the multiple regression output (see Appendix B1) shows the adjusted R^2 of .2545 which means that the model explains around 25.5% of the variance of the target variable which is slightly under .26, with .26 considered as a high effect of the model (Field, 2013). The model is significant at a $p < .001$ level. The intercept is also significant at a $p < .001$ level, however further interpretation of the intercept would not make sense. This is due to the fact that no post can have all the predictor values at zero since year, month, weekday and type are dummy variables without a zero value.

5.2.3.2 Overall fit of the predictors

An analysis of variance (ANOVA) output shows the overall fit of the predictors in the model where the most interesting value is the F-ratio (column Pr(>F)) showing whether the results of the model are by chance (Field, 2013). Table 5.7 shows that all the predictors have a highly significant effect on the model since the column Pr(>F) has the majority of values under .001 with the exception of weekday which is significant at a $p < .05$ level. However, the results of the ANOVA do not explain the individual impact of predictors in the model (Field, 2013). Thus, even though the predictors all show a high significance level in the overall model, this is not sufficient to confirm or reject the hypotheses.

Table 5.7 ANOVA

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(> F)**
Type	4	17.8	4.5	94.6	0
City	1	0.8	0.8	16.2	0.0001
Length	1	4.6	4.6	98.4	0
Month	11	3.7	0.3	7.2	0
Hour	23	9.4	0.4	8.7	0
Year	7	31	4.4	94	0
Weekday	6	0.8	0.1	2.7	0.014
Residuals	4,029	189.6	0		

The full results of the multiple regression model are shown in Appendix B1 where the baseline dummy variable group for each predictor is automatically excluded from the regression result.

R automatically chooses the baseline to be the first category in the group either by numeric order, starting from hour 0, month 1, city 0 and year 2009, or by alphabetic order, starting from type event and weekday Friday. As explained in section 3.2.3, simple linear regression models are used to explain the specific impacts of predictors on the target. The following section presents the results of the simple linear regression models.

5.2.4 Simple linear regression models

The results of the simple linear regressions can be found in Appendix B2. In the following sections, the correlation values between variables are reported, including the significance level. The rounded mean, reported only for the baseline, is denoted with β_0 and β (beta) stands for the difference of the dummy variable mean and the baseline mean. An example of the interpretation is as follows. The baseline estimate for type is 0.97. Since “type = event” is the only type not seen in the other rows of the model summary, it can be inferred that “type = event” is the baseline value. The value for the baseline is denoted as follows: $\beta_0 = 0.97$, $p < .001$. The estimate for “type = link”, for example, is -0.17, which is denoted as $\beta = -0.17$, $p < .01$.

Since the target value is scaled to a range between 0 and 1, the estimate values what proportion of the maximum possible target value is expected for a certain predictor value. For example, for “type = link” the expected value is 0.8 (0.97 – 0.17) out of 1, meaning the average expected value of the target is 80% of the maximum included value for the target when “type = link”. As stated before, there is no positive or negative direction associated with any of the categorical predictors. Instead, the directions associated with the estimates will be benchmarked against the estimate of the baseline value. In short, following Field (2013) β stands for the unstandardized beta value that shows the relative difference of the specific dummy variable and the baseline group. The β value represents the change in the target "*due to a unit change in the predictor*" (Field, 2013, p. 424).

As can be deducted from the results displayed in Appendix B2, the residuals of all individual simple linear regression models are normally distributed as the first and third quartiles are within a ±1.5 range of the residual standard error, indicating that linear regression is a suitable method for each case. The overall average value of the target variable in the dataset used for the regression models is 0.728 out of 1; this value plays an essential role as a benchmark for the direction of the effects caused by each predictor and their possible values. For each

categorical predictor, the expected mean value for the target variable will be presented in a table, accompanied by the overall value of the target variable (0.728) and the difference between the predictor value and the target variable value (see Appendix C, line 214-220). This provides insight in how each value differs from the overall average, and whether a certain value for a predictor has a positive or negative effect on the target variable compared to the baseline; whether this effect is significant, is then inferred from the linear regression models. These tables should be seen as a complementary tool to enable easy interpretation of the linear regression model outputs, rather than an integral part of the regression models.

5.2.4.1 Effect of city

The simple linear regression model for the city or user account that a post was created by (VisitCopenhagen or VisitTallinn) significantly affects the target variable. Posts created by VisitCopenhagen receive an average score for the target of $\beta_0 = 0.744$ out of 1. Posts created by VisitTallinn perform relatively worse compared to the baseline, as the estimate for this city is 0.704 (β_0 0.744 – β 0.040), $p < .001$. The difference is small, however. This implication does not relate to any hypothesis, but it does provide a comparison of the two cities considered. The deviance from the overall mean is rather low, so it is not surprising that this predictor has a small explanatory value when it comes to explaining the variance in the target variable (adjusted $R^2 = 0.057$).

Table 5.8 Mean target values city

City	Mean target value	Overall mean	Difference	β	Sig.
0 (Copenhagen)	0.744	0.728	0.016		< .001 ***
1 (Tallinn)	0.704	0.728	-0.024	-0.04	< .001 ***
Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05					

5.2.4.2 Effect of post type

Regarding the type of posts, since β of "type = link" has a negative value, $\beta = -0.174$, $p < .01$, it shows that the target value decreases as a post changes from the baseline type event, $\beta_0 = 0.975$, $p < .001$, to type link. As Field (2013) explains, the relative decrease of β value compared to the baseline represents a bigger change, i.e. a post receives less likes, comments

and/or shares so this actually shows that the target value decreased significantly more in type link posts compared to type event posts. The same counts for the other types; photo and video. Type video is estimated to relatively decrease the target value compared to the baseline, $\beta = -0.193$, $p < .01$, followed by type photo, $\beta = -0.299$, $p < .001$. Thus, H1 is partially supported as most types of posts have a statistically significant influence on the target.

Table 5.9 Mean target values type

Type	Mean target value	Overall mean	Difference	β	Sig.
event	0.975	0.728	0.246		< .001 ***
link	0.801	0.728	0.073	-0.174	.008 **
photo	0.676	0.728	-0.052	-0.298	< .001 ***
status	0.861	0.728	0.133	-0.114	.09
video	0.781	0.728	0.053	-0.193	.004 **

Signif. codes: '***' .001 '**' .01 '*' .05

5.2.4.3 Effect of length

The predictor length has a highly statistically significant effect on the target, $\beta = 0.001$, $p < .001$, supporting H2. The direction of the effect is positive, although the β value is close to 0, meaning that one additional word leads to only slightly higher values of the target variable. The intercept is significant, $\beta_0 = 0.702$, $p < .001$, showing that when the post includes no words, meaning $X = 0$, the regression model predicts that a post will receive 92 likes (Field, 2013). More specifically, the results of the simple regression analysis concerning post length shows that when the target is scaled to a range between 0 and 1, then one additional word per post causes an increase of 0.077 (β 0.00077 * 100) percentage points on the target. As this effect is accomplished by merely one extra word, the magnitude of the effect is expected to be much larger when a post is prolonged by several words or even sentences. These results support H2 confirming that text length statistically significantly influences the amount of likes, comments and shares that a post receives.

5.2.4.4 Effect of year

Posts published in the baseline year 2009 have a significant positive impact on the target, β_0 = 0.881, p < .001. The following years show a significant negative trend in the target, from 2012, β = -0.064, p < .01, until 2016, β = -0.38, p < .001. Interestingly, the relative decrease of the positive effect of the year a post was created grows gradually compared to the year before, indicating that as years pass, posts have been receiving fewer likes, comments and/or shares. This can be explained by the fact that the in following years after 2009 there has been an increase in different social media platforms. As tourists have more options in choosing where to keep themselves up-to-date with tourism related information, the amount of reactions that one specific page receives decreases. The adjusted R^2 of the simple linear regression concerning year of creation is relatively high as compared to those of the other simple linear regression models (0.195), indicating that this predictor explain much of the variance in the overall multiple linear regression model with respect to the other predictors. Moreover, the majority of the estimates is highly significant, further contributing to the explanatory value of the model. The results of this model support H3.

Table 5.10 Mean target values year

Year	Mean target value	Overall mean	Difference	β	Sig.
2009	0.881	0.728	0.153		< .001 ***
2010	0.873	0.728	0.144	-0.008	.753
2011	0.868	0.728	0.140	-0.013	.598
2012	0.817	0.728	0.089	-0.064	.006 **
2013	0.744	0.728	0.016	-0.137	< .001 ***
2014	0.656	0.728	-0.072	-0.225	< .001 ***
2015	0.637	0.728	-0.091	-0.243	< .001 ***
2016	0.501	0.728	-0.227	-0.38	< .001 ***
Signif. codes: '***' .001 '**' .01 '*' .05					

5.2.4.5 Effect of month

The simple linear regression model with month of creation as a predictor shows varying results in terms of significance. In particular, posts published in the baseline month January (month01)

have a significant positive impact on the target, $\beta_0 = 0.711$, $p < .001$. August (month08) shows the largest estimate, $\beta = 0.054$ which is the most significant result after the baseline at a significance level of $p < .01$. July (month07) and December (month12) show relatively high estimates as well, at $\beta = 0.044$ and $\beta = 0.047$ respectively, with a significance level of $p < .05$. It can thus be inferred that posting in July, August and December positively affect the target variable, as compared to the baseline value of January. This could be explained for instance by the fact that both of the month pairs, July, August and December, January, include school holiday periods in both Denmark and Estonia where more events are taking place. Contrastingly, March (month03) and April (month04) show a slight negative impact compared to the baseline with $\beta = -0.036$ and $\beta = -0.043$ respectively, with a significance level of $p < .05$. Since the effect of the month of creation was shown to be significant for half of the months, H4 is partially supported.

Table 5.11 Mean target values month

Month	Mean target value	Overall mean	Difference	β	Sig.
01	0.711	0.728	-0.017		< .001 ***
02	0.710	0.728	-0.018	-0.001	.942
03	0.675	0.728	-0.053	-0.036	.048 *
04	0.668	0.728	-0.060	-0.043	.021 *
05	0.747	0.728	0.019	0.036	.061
06	0.749	0.728	0.021	0.038	.058
07	0.755	0.728	0.027	0.043	.026 *
08	0.765	0.728	0.037	0.054	.004 **
09	0.740	0.728	0.012	0.029	.13
10	0.744	0.728	0.016	0.033	.081
11	0.731	0.728	0.003	0.02	.292
12	0.758	0.728	0.030	0.047	.013 *
Signif. codes: '***' .001 '**' .01 '*' .05					

5.2.4.6 Effect of weekday

The model targeting weekday only shows a significant outcome for Friday – the baseline – and Sunday, at $\beta_0 = 0.737$ and $\beta = -0.063$ respectively, with a significance level of $p < .001$. The effect for Sunday is rather strong, indicating that posting on a Sunday has a considerable negative impact on the target when compared to Friday. Overall, H5 is partially supported.

Table 5.12 Mean target values weekday

Weekday	Mean target value	Overall mean	Difference	β	Sig.
Friday	0.737	0.728	0.008		< .001 ***
Monday	0.734	0.728	0.006	-0.003	.84
Saturday	0.722	0.728	-0.006	-0.015	.344
Sunday	0.674	0.728	-0.054	-0.063	< .001 ***
Thursday	0.732	0.728	0.003	-0.005	.711
Tuesday	0.721	0.728	-0.007	-0.016	.256
Wednesday	0.759	0.728	0.031	0.022	.105

Signif. codes: '***' .001 '**' .01 '*' .05

5.2.4.7 Effect of hour

Post created between 3 AM and 4 AM receive significantly fewer likes/comments/shares, $\beta = -0.278$, $p < .05$ than posts created between 12 AM and 1 AM (the baseline), $\beta_0 = 0.753$ with a significance level of $p < .001$. The significance of these results is not very strong, however, and other hours of creation do not have any significant effect. Thus, H6 is rejected.

Table 5.13 Mean target values hour

Hour	Mean target value	Overall mean	Difference	β	Sig.
00	0.753	0.728	0.025		< .001 ***
01	0.554	0.728	-0.174	-0.199	.093
02	0.780	0.728	0.052	0.027	.837
03	0.475	0.728	-0.253	-0.278	.026 *
04	0.600	0.728	-0.128	-0.153	.167

05	0.725	0.728	-0.004	-0.028	.717
06	0.760	0.728	0.032	0.007	.925
07	0.790	0.728	0.062	0.037	.627
08	0.787	0.728	0.059	0.034	.65
09	0.802	0.728	0.073	0.049	.521
10	0.797	0.728	0.069	0.044	.565
11	0.757	0.728	0.029	0.004	.953
12	0.765	0.728	0.037	0.012	.873
13	0.767	0.728	0.038	0.014	.858
14	0.712	0.728	-0.016	-0.041	.589
15	0.648	0.728	-0.080	-0.105	.164
16	0.678	0.728	-0.050	-0.075	.32
17	0.680	0.728	-0.048	-0.073	.333
18	0.666	0.728	-0.062	-0.087	.249
19	0.647	0.728	-0.081	-0.106	.163
20	0.703	0.728	-0.025	-0.05	.519
21	0.818	0.728	0.090	0.065	.422
22	0.741	0.728	0.013	-0.012	.89
23	0.786	0.728	0.058	0.033	.717
Signif. codes: '***' .001 '**' .01 '*' .05					

5.2.4.8 Effect of sentiment and intensity

The results of the simple linear regression models regarding the tone and intensity of posts can be found in Appendix B3, while the corresponding R code can be found in Appendix C (line 222-233). As the targets – likes, comments and shares – are not correlated (see Table 5.3), separate regressions are run for the three targets and the two predictors, resulting in six simple linear regression models in total. Again, the target variables are scaled to a range between 0 and 1 for the purpose of interpretation.

5.2.4.8.1 Sentiment

Regarding the predictive power of the simple regression models the adjusted R-squared of each model predicting sentiment is low, ranging from -0.015 in the case of likes to 0.103 in the case of comments. This can partially be explained by the low N (N = 215). A very low or negative adjusted R-squared shows that the model does not significantly predict the target value and the results are by chance (Field, 2013).

The number of likes a post receives is significantly affected by its tone, as a more positive tone has a relative negative impact on the number of likes with $\beta = -0.044$, $p < .05$ compared to the baseline, neutral tone, $\beta_0 = 0.115$, $p < .001$.

Table 5.14 Mean target values sentiment and likes

Sentiment	Mean target value	Overall likes mean	Difference	β	Sig.
Neutral	0.115	0.078	0.037		< .001 ***
Positive	0.071	0.078	0.007	-0.044	.041 *
Signif. codes: '***' .001 '**' .01 '*' .05					

Similar results apply to the effect of positive tone on the number of comments; this is a rather strong and significant negative effect with $\beta = -0.1$, $p < .001$ compared to the baseline, neutral, $\beta_0 = 0.169$, $p < .001$.

Table 5.15 Mean target values sentiment and comments

Sentiment	Mean target value	Overall comments mean	Difference	β	Sig.
Neutral	0.169	0.084	0.085		< .001 ***
Positive	0.069	0.084	-0.015	-0.1	< .001 ***
Signif. codes: '***' .001 '**' .01 '*' .05					

Posts with a neutral tone significantly affect the amount of shares a post receives, $\beta_0 = 0.05$, $p < .05$. The influence of positive tone on the shares count a post receives is not statistically significant.

Table 5.16 Mean target values sentiment and shares

Sentiment	Mean target value	Overall shares mean	Difference	β	Sig.
Neutral	0.05	0.032	0.018		.027 *
Positive	-0.821	0.032	-0.853	-0.871	.385
Signif. codes: '***' .001 '**' .01 '*' .05					

Overall, posts with a positive tone are expected to receive significantly fewer likes and comments than posts with a neutral tone. Thus, H7 is supported.

5.2.4.8.2 Intensity

Posts with low intensity (1) have a significant influence on the likes count a post receives, $\beta_0 = 0.085$, $p < .001$. The influence of medium (2) and high (3) intensity posts on the likes count a post receives is not statistically significant.

Table 5.17 Mean target values intensity and likes

Intensity	Mean target value	Overall likes mean	Difference	β	Sig.
1	0.085	0.078	0.007		< .001 ***
2	0.071	0.078	-0.007	-0.014	.538
3	0.084	0.078	0.006	-0.001	.966
Signif. codes: '***' .001 '**' .01 '*' .05					

A post's intensity significantly affects the number of comments a post receives. Interestingly, a stronger intensity negatively affects the number of comments compared to the baseline, $\beta_0 = 0.14$, $p < .001$. Namely, with $\beta = -0.062$, $p < .01$ for an intensity of 2 and with $\beta = -0.072$, $p < .01$ for an intensity of 3.

Table 5.18 Mean target values intensity and comments

Intensity	Mean target value	Overall comments mean	Difference	β	Sig.
1	0.14	0.084	0.056		< .001 ***

2	0.078	0.084	-0.006	-0.062	0.004 **
3	0.068	0.084	-0.016	-0.072	0.001 **

Signif. codes: '***' .001 '**' .01 '*' .05

The influence of post's intensity is not statistically significant on the amount of shares a post receives. Thus, H8 is partially supported.

Table 5.19 Mean target values intensity and shares

Intensity	Mean target value	Overall shares mean	Difference	β	Sig.
1	0.018	0.032	-0.014		.398
2	0025	0.032	-0.007	0.007	.786
3	0.046	0.032	0.014	0.028	.279

Signif. codes: '***' .001 '**' .01 '*' .05

5.3 Comparison between Copenhagen and Tallinn

In order to compare the effects of the predictors on the target value, separate analyses were run for Copenhagen and Tallinn (see Appendix C, line 235-259). Two additional multiple linear regression models were run after selecting all posts from either Copenhagen or Tallinn, followed by PCA, removing outliers and normalizing the target variable. The results can be found in Appendix B4.

Interestingly, the type of a post has a significant effect for every post type considered in the case of Tallinn, while the post type is not significant in any case for Copenhagen. In particular, posts that are labeled as photos have a particularly relative negative effect on the target variable when compared to the baseline, with $\beta = -0.279$, $p < .001$. The direction and strength of this effect is in line with the results of the general multiple linear regression model (see Appendix B1). These results suggest that for Copenhagen, the type of post is irrelevant when it comes to the gathered number of likes, comments and/or shares, while for Tallinn it is highly relevant. This can partially be explained by the fact that Tallinn is still a developing tourism destination and promotion of the city has a high priority (Tooman & Müristaja, 2014). Thus, it is important

which types of posts are published on the city branding page as the visual content of posts, for instance in the case of videos, can greatly promote the city's image.

Table 5.20 Multiple regression statistically significant results for Copenhagen

Predictor	Mean target value	Overall mean	Difference	β	Sig.
Intercept	1.111	0.744	0.367		< .001 ***
length	1.112	0.744	0.368	0.001	< .001 ***
month04	1.067	0.744	0.323	-0.044	.032 *
2013	1.013	0.744	0.269	-0.098	.002 **
2014	0.83	0.744	0.086	-0.281	< .001 ***
2015	0.811	0.744	0.067	-0.3	< .001 ***
2016	0.71	0.744	-0.034	-0.401	< .001 ***
Tuesday	1.076	0.744	0.332	-0.035	.021 *
Signif. codes: '***' .001 '**' .01 '*' .05					

For the month of a post, no noteworthy differences exist, although the significance of the months differs for Copenhagen (April has a significant effect) and Tallinn (November and December have a significant effect). For the hour of a post, no results for Copenhagen nor Tallinn are statistically significant. When comparing the year a post was a created, the results are statistically significant for Copenhagen between years 2013 until 2016 and for Tallinn for 2016. The effect of posting in 2016 is relatively strong and negative in both cases compared to the baseline, while the negative effect is stronger for Copenhagen. The influence of the years from 2013 until 2015 is negative on the amount of likes, comments and/or shares a post receives compared to the baseline in the case of Copenhagen. Thus, the negative effect, compared to the baseline, of posting in recent years is stronger as well as more significant for Copenhagen than for Tallinn.

Table 5.21 Multiple regression statistically significant results for Tallinn

Predictor	Mean target value	Overall mean	Difference	β	Sig.
Intercept	0.798	0.717	0.081		< .001 ***
link	0.628	0.717	-0.089	-0.17	.005 **

photo	0.519	0.717	-0.198	-0.279	< .001 ***
status	0.616	0.717	-0.101	-0.182	.007 **
video	0.603	0.717	-0.114	-0.195	.003 **
length	0.799	0.717	0.082	0.001	< .001 ***
month11	0.73	0.717	0.013	-0.068	.015 *
month12	0.731	0.717	0.014	-0.067	.015 *
2016	0.526	0.717	-0.191	-0.272	< .001 ***

Signif. codes: '***' .001 '**' .01 '*' .05

In regard to the weekday, only the model for Copenhagen had statistically significant results for Tuesday, $\beta = -0.035$, $p < .05$. The length is statistically significant for both Tallinn, $\beta = 1.112$, $p < .001$, and Copenhagen, $\beta = 0.799$, $p < .001$. All in all, the separate multiple regression models run for Copenhagen and Tallinn show certain communalities, although the majority of the results are different for the two cities. This can be partially explained by the different marketing strategies of the cities. Copenhagen is a popular tourist destination where a larger number of people are visiting the page since more tourists visit the city compared to Tallinn. However, Tallinn is a developing Eastern European capital that is still exploring the opportunities of city branding.

5.4 Manual sentiment analysis

This section explains the results of the manual sentiment analysis. The manual analysis of 215 posts provides insights on the tones, topics and intensities of the selected posts. The posts are imported from R to an Excel file where the analysis is carried out. For the descriptive analysis the results were imported back to R. A sample of the manually analyzed data can be found in Appendix A2, while the corresponding R code can be found in Appendix C (line 198-211).

The underlying tone of a post sheds light on the sentiment of a post at a general level. Regarding the tone, 33 posts are considered neutral and 182 posts are considered positive. This shows a high frequency of positive posts. This can be explained by the trend that these Facebook pages are visited by a community of people who have an interest in a city and generally are there to share positive experiences and views. Furthermore, some of the neutral

posts were not related to the page or the city itself but, for instance were sympathetic towards Norway due to the 2011 attacks.

The magnitude of the sentiment is measured by intensity, which adds an additional layer to the analysis. As for the intensity, 35 posts have an intensity score of 1, 102 posts have an intensity score of 2 and 78 posts have an intensity score of 3. This shows a rather normal distribution, with a slight lean towards a high intensity. The mode, as well as median, value is 2 on a range of 1 until 3, ranging from low to high. Furthermore, 45 posts (21%) include one or more emoticons, while 98 posts (46%) have an exclamation mark in them. 188 posts (87%) do not include words in all capital letters, while 27 posts (13%) do.

Figure 5.13 shows the distribution of the emotions addressing the posts in a more detailed manner. It is evident that "friendly" and "anticipation" are by far the most commonly used emotions in the posts, while "gratitude", "loving" and "dispirited" are not found very often. The latter is not surprising, since such a sentiment is intuitively associated with a negative tone, and only one post has a negative tone. "Anticipation" was often recognized for posts that were made prior to an upcoming cultural activity or an event. "Friendly" was also often used in regard to an event but more in a welcoming sense where information was given about the location or artists. "Pride" was often used relating to athletes and sports events.

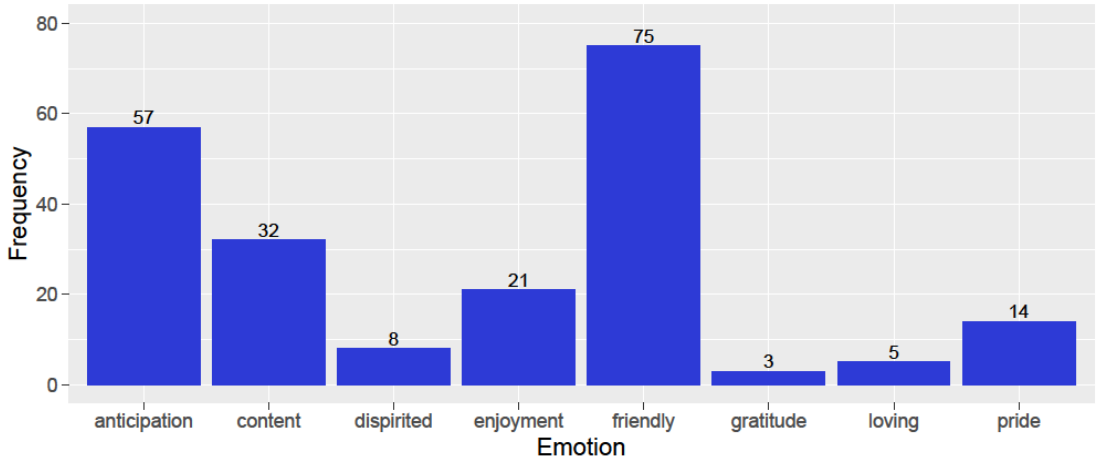

Figure 5.13 Frequencies of sentiments

The distribution of the assigned codes are displayed in Table 5.22. By far most posts are about culture (69 posts; 32% of the total), followed by events (33 posts; 15% of the total) and tourism (32 posts; 15% of the total). These results emphasize the expected purpose of the Facebook accounts, namely to target tourists and to share information on events that are happening in the

city as well as cultural information. Almost all the posts are assigned a meaningful code, except one post that was assigned "other". The post included the message "*Is today a day like any other?*" which does not allow any categorization in regard to an underlying theme. The "news" code does not mean that there were actual news company posts present on the page, the messages in this category included also an example like "*has some exciting news, but can't tell until Monday :D Have a great weekend in suspension everyone!"* This is due to the fact that the aim of the coding was to categorize as many posts as possible and too strict parameters for a specific code would in the end either create too many different codes or leave many posts under "other".

Table 5.22 Frequencies of manual codes

Code	Frequency	Code	Frequency
culture	69	food	5
event	33	weekend	5
tourism	32	shopping	4
sports	21	drink	3
news	15	city tour	1
weather	14	transport	1
holidays	11	other	1

The manual sentiment analysis proved to be a rich supplement to the quantitative research of the study as it offered a more detailed look at the data. The results of sentiment analysis were according to expectations and there were no shocking occurrences or extreme outliers present. All in all, the underlying sentiment of the posts is positive and handles mainly topics related to the cultural sphere as well as tourism.

6 Conclusions and implications

The aim of this study is to explore the relationships between local governmental branding posts and eWOM reactions, in the form of likes, comments and shares, of tourists. Conducting analysis on a sample of 4,083 posts that were published on the VisitTallinn and VisitCopenhagen page, this study provides insight into the different predictors that influence the reaction a posts receives. This study shows that both length of the post, as well as type and time of posting influence the amount of likes, comments and/or shares the post receives. Various differences and commonalities between Copenhagen and Tallinn were exposed. Furthermore, the manual sentiment analysis gives an enriched insight into the underlying emotions of the reactions received.

The main research question of this study is:

> *What is the influence of local governmental city branding on eWOM communication by tourists in Copenhagen and Tallinn?*

In order to answer the central research question, multiple and simple linear regression analyses were conducted to test how six independent variables are able to predict the target, namely the number of likes, comments and shares that posts receive from tourists. There are several statistically significant variables that have an impact on this target. The highest impact was proven to be on the type of posts, as well as length of posts. The year of the posts also proved to be highly significant. A positive tone and a high intensity were surprisingly shown to negatively affect likes, comments or shares.

6.1 Discussion

Brand managers can use the findings of this study to adjust their marketing strategies to induce an active response from tourists. This study reveals the factors of a post that have to be considered in order to induce user engagement. Identifying these factors allows brand managers to pursue a more user-oriented marketing strategy by focusing on the factors that impact user responsiveness (Hennig-Thurau et al., 2004). Through analyzing the reactions of tourists, local government can investigate the tourists' needs and possibly involve tourists in co-creation by inviting them to express their opinions and ideas (Cvijikj & Michahelles, 2013).

This type of co-creation strengthens the brand in the sense that there is less chance for opposition since by including tourists in the brand development, tourists bear some of the responsibility for the brand.

This study can partially confirm the findings of Cvijikj et al. (2011) showing that different types of posts have a significant effect, however status type posts did not have a significant effect on the eWOM communication by tourists. The latter can be explained by the fact that proportionally there is a small amount of status type posts in the data. After 2010 the amount of status posts started to decrease drastically resulting in no status type posts occurring after 2014. The results of this study show that photos received the least amount of likes, comments and/or shares compared to other types of posts, which on one hand is not in line with the findings of Malhotra et al. (2013), who found photos to be the most popular type of posts. On the other hand, the findings of this study confirm the results of Cvijikj et al. (2011) who found photos on brand pages to induce the least reactions from users. The results of this study can partially be explained by the fact that photos include merely illustrated content and receive lower attention compared to videos, which have richer content, and links that redirect tourists to further information (Cvijikj & Michahelles, 2013).

The results of this study supporting H2 are in line with the research of Sabate et al. (2014) who find a positive impact of the text length on the amount of likes. The average text length of a post is 33 words, which is slightly more than de Vries et al. (2012) found, namely 28 words, researching brand banners, i.e. advertisements that are aimed to induce people to click on them. This can be explained by the fact that while banners primarily aim to attract attention quickly (De Vries et al., 2012), local government brand posts also aim to inform and engage tourists on a longer term, for instance by showcasing the unique elements of a city. This finding is contradictory to Malhotra et al. (2013) who advise to keep posts as brief as possible in order to receive more likes.

In regard to the weekday only posting on Friday and Sunday revealed statistically significant results. This shows that in general posting on different weekdays does not have a strong influence on the amount of likes, comments and/or shares a post receives. Therefore, posting daily seizes opportunities to receive eWOM reactions from tourists (Carter, 2014). This is also a point that can be difficult to include in a marketing strategy since when an event or update is announced, it is not intelligible to delay posting about it. Since when a page is not up-to-date tourists begin to explore other sources for actual information.

While factors such as the length, type, tone, topic and date and time of creation of posts generated by government city brand managers targeting tourists alone are not decisive factors in a city's image among tourists, these aspects of social media communication do have a place in a city's overlapping brand management. As such, the results of this study allow city brand managers to further steer their marketing strategy and improve their communication channels with tourists. More elaborate recommendations are presented in chapter 7.

6.2 Assumptions and limitations

This study's research design, data collection method and data analysis method are subject to a number of assumptions and exhibit several limitations.

First, while it cannot be determined with certainty that people reacting to local governmental city branding posts on the selected pages are tourists, it is assumed that this is the case since the selected pages specifically target tourists. Due to Facebook privacy restrictions it was not possible to conduct a user-based analysis.

Second, the nature of retrospective data collection is prone to outliers since some data might have been lost or not documented (Montgomery, Peck, & Vining, 2006). Third, even though quantitative analysis was supplemented with sentiment analysis, the majority of user input was not researched from the content side. For instance, the topic, tone and intensity of user comments has not been taken into account in the analysis of the influence of branding activities on eWOM communication by tourists. Fourth, potential mediating and moderating factors have not been taken into account, partially due to lack of such factors being grounded in theory, and partially due to lack of further data.

6.3 Future research

First, future research can include a wider spectrum of countries to identify trends on a broader scale, and possibly to further delve into cultural or regional differences. Second, for the sentiment analysis a greater sample size would reveal more detailed analysis. With a larger sample size, the results of the linear regression models concerning sentiment may become more powerful in terms of explanatory value. Third, advanced automatic sentiment analyses tools

could be applied, for example using R. This would allow larger sets of data to be analyzed, targeting sentiment, topic and intensity.

Fourth, it would be interesting to conduct a similar research based on user input, meaning the content of user comments could be researched, as well as a survey study could be conducted. A survey can also reveal more about the user profiles who are reacting to the posts, which would add an additional dimension to the study. For instance, future research could explore whether an age gap, identified by Bimber (1999), persists due to differences in familiarity towards the Internet.

Fifth, this study can be replicated on different brand pages which allows comparing the results between different local government sectors, such as education or health care, as well as the private sector. Sixth, the methods of this study could be supplemented with qualitative methods, such as interviews with the city brand managers organizing the Facebook pages as well as interviews with tourists and users visiting the pages. This would provide insight into the reasons behind why and how the people on both sides use these social media communication channels. Seventh, with regard to one of the limitations mentioned in the previous section, future research may include moderating and mediating variables, which might possibly uncover their role.

7 Recommendations

Local government city brand managers can use the findings of this study to develop marketing strategies that improve the attractiveness of a city for tourists. First, the overall trend, since 2014, of increasing popularity of posting on Facebook exemplifies the importance of staying active on Facebook. While Facebook popularity gradually decreased between 2012 and 2014, the latest years are showing that Facebook is regaining momentum. In order to enhance the reactions a post receives on Facebook city brand managers can investigate the needs and interests of current and potential tourists. This can be done for instance through online surveys or by posting questions, where tourists are invited to express their opinion about a topic. In particular, the city brand managers of the Copenhagen social media account ought to address the decrease in the number of likes, comments and/or shares, since the negative impact of recent years has been stronger for Copenhagen than for Tallinn.

Furthermore, city brand managers need to address what type of posts attract likes, comments and/or shares. For city brand managers of the Tallinn social media account in particular, attention needs to be paid to why certain types of posts negatively affect the number of likes, comments and/or shares received. Copenhagen's city brand managers will have to perform additional research, as in their case the post type does not affect the number of likes, comments and/or shares at all. Their marketing research ought to figure out why this is the case, and how this can be changed so that their social media success and reach can be steered by the type of content they post.

The results of this study further show that more positive and more intense posts negatively affect the number of likes, comments or shares received, which may be counterintuitive. City brand managers can consider this, since as a reasonable yet inaccurate assumption one might make is that positive and intense posts extend the reach and popularity of a post. The opposite is shown to be true, and marketing strategies can be adapted accordingly.

By identifying topics that are relevant for tourists brand managers are able to post about issues that intrigue tourists. This is what Malhotra et al. (2013) call being topical. For instance, when a popular music festival is approaching the city brand page should include information about the event with subtle hints to the city brand (Malhotra et al., 2013). This creates a chance for people to, for instance share their feeling of anticipation with others through eWOM. Furthermore, this links the brand with a popular event promoting the city with the support of

a renowned brand. This is not a new tool in the marketing strategy toolbox (Braun, 2011), however it can be further incorporated in local government city branding.

References

Barbera, P. (2016). Rfacebook: Access to Facebook API via R. R package version 0.6.3. Retrieved from https://cran.r-project.org/web/packages/Rfacebook/Rfacebook.pdf

Bimber, B. (1999). The Internet and citizen communication with government: Does the medium matter? *Political communication, 16*(4), 409-428. doi:10.1080/105846099198569

Bonsón, E., Torres, L., Royo, S., & Flores, F. (2012). Local e-government 2.0: Social media and corporate transparency in municipalities. *Government information quarterly, 29*(2), 123-132. doi:10.1016/j.giq.2011.10.001

Braun, E. (2011). Putting city branding into practice. *Journal of Brand Management, 19*(4), 257-267. doi:10.1057/bm.2011.55

Bryer, T. A., & Zavattaro, S. M. (2011). Social media and public administration: Theoretical dimensions and introduction to the symposium. *Administrative Theory & Praxis, 33*(3), 325-340. doi:10.2753/ATP1084-1806330301

Carter, B. (2014). *The Like Economy: How Businesses Make Money with Facebook* (2nd ed.). Indianapolis: Que Publishing.

Chadwick, A. (2006). *Internet politics: States, citizens, and new communication technologies.* New York: Oxford University Press.

Chan, K. W., & Li, S. Y. (2010). Understanding consumer-to-consumer interactions in virtual communities: The salience of reciprocity. *Journal of Business Research, 63*(9), 1033-1040. doi:10.1016/j.jbusres.2008.08.009

Cohen, J. (1992). A power primer. *Psychological bulletin, 112*(1), 155-159.

Cvijikj, I. P., & Michahelles, F. (2013). Online engagement factors on Facebook brand pages. *Social Network Analysis and Mining, 3*(4), 843-861. doi:10.1007/s13278-013-0098-8

Cvijikj, I. P., Spiegler, E. D., & Michahelles, F. (2011). The effect of post type, category and posting day on user interaction level on Facebook. *Privacy, Security, Risk and Trust (PASSAT)* (pp. 810-813). IEEE Third Inernational Conference on Social Computing (SocialCom).

Dahlgren, P. (2005). The Internet, public spheres, and political communication: Dispersion and deliberation. *Political communication, 22*(2), 147-162. doi:10.1080/10584600590933160

Dahlgren, P. (2013). *The political web: Media, participation and alternative democracy.* Basingstoke: Palgrave Macmillan. doi:10.1057/9781137326386

De Vries, L., Gensler, S., & Leeflang, P. S. (2012). Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing. *Journal of Interactive Marketing, 26*(2), 83-91. doi:10.1016/j.intmar.2012.01.003

Duggan, M., Ellison, N. B., Lampe, C., Lenhart, A., & Madden, M. (2015). *Social media update 2014.* Washington, D.C: Pew Research Center.

Durbin, J., & Watson, G. S. (1951). Testing for serial correlation in least squares regression. II. . *Biometrika, 38*(1/2), 159-177.

Eshuis, J., & Klijn, E. H. (2012). *Branding in governance and public management.* London: Routledge.

Eshuis, J., Klijn, E. H., & Braun, E. (2014). Place marketing and citizen participation: branding as strategy to address the emotional dimension of policy making? *International Review of Administrative Sciences, 80*(1), 151-171. doi:10.1177/0020852313513872

Eurostat. (2015, December 22). *Individuals - internet use.* Retrieved from Eurostat: http://appsso.eurostat.ec.europa.eu/nui/show.do

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of statistical software, 25*(5), 1-54.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (4th ed.). London: Sage.

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative inquiry, 12*(2), 219-245. doi:10.1177/1077800405284363

Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association, 87*(417), 178-183.

Gerbner, G. (1969). Toward "cultural indicators": The analysis of mass mediated public message systems. *Educational Technology Research and Development, 17*(2), 137-148.

Hankinson, G. (2001). Location branding: A study of the branding practices of 12 English cities. *The Journal of Brand Management, 9*(2), 127-142. doi:10.1057/palgrave.bm.2540060

Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of interactive marketing, 18*(1), 38-52. doi:10.1002/dir.10073

Hlavac, M. (2015). stargazer: Well-formatted regression and summary statistics tables. R package version 5.2. Retrieved from https://cran.r-project.org/web/packages/stargazer/stargazer.pdf

Jørgensen, L. B., & Munar, A. M. (2009). Chapter 13 THE COPENHAGEN WAY Stakeholder-driven Destination Branding. In A. C. Liping, W. C. Gartner, & A. M. Munar (Eds.), *Tourism Branding: Communities in Action* (pp. 177-189). Bradford: Emerald Group Publishing Limited. doi:10.1108/S2042-1443(2009)0000001015

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement, 20*(1), 141-151. doi:10.1177/001316446002000116

Kavanaugh, A. L., Fox, E. A., Sheetz, S. D., Yang, S., Li, L. T., Shoemaker, D. J., . . . Xie, L. (2012). Social media use by government: From the routine to the critical. *Government Information Quarterly, 29*(4), 480-491. doi:10.1016/j.giq.2012.06.002

Kavaratzis, M. (2004). From city marketing to city branding: Towards a theoretical framework for developing city brands. *Place Branding, 1*(1), 58-73. doi:10.1057/palgrave.pb.5990005

Kavaratzis, M., & Ashworth, G. J. (2005). City branding: An effective assertion of identity or a transitory marketing trick? *Tijdschrift Voor Economische En Sociale Geografie, 96*(5), 506-514. doi:10.1111/j.1467-9663.2005.00482.x

Kim, B. (2015). Understanding diagnostic plots for linear regression analysis. Retrieved from http://data.library.virginia.edu/diagnostic-plots/

Klijn, E. H., Eshuis, J., & Braun, E. (2012). The influence of stakeholder involvement on the effectiveness of place branding. *Public Management Review, 14*(4), 499-519. doi:10.1080/14719037.2011.649972

Klinger, U., & Svensson, J. (2014). The emergence of network media logic in political communication: A theoretical approach. *New Media & Society, 17*(8), 1241-1257. doi:10.1177/1461444814522952

Kwok, L., & Yu, B. (2013). Spreading social media messages on Facebook: An analysis of restaurant business-to-consumer communications. *Cornell Hospitality Quarterly, 54*(1), 84-94. doi:10.1177/1938965512458360

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological bulletin, 68*(5), 304-305.

Louw, P. E. (2005). *The Media and Political Process.* London: SAGE.

Lowndes, V., & Roberts, M. (2013). *Why institutions matter: The new institutionalism in political science.* Basingstoke: Palgrave Macmillan.

Malhotra, A., Malhotra, C. K., & See, A. (2013). How to create brand engagement on Facebook. *MIT Sloan Management Review, 54*(2), 17-20.

Merz, M. A., He, Y., & Vargo, S. L. (2009). The evolving brand logic: a service-dominant logic perspective. *Journal of the Academy of Marketing Science, 37*(3), 328-344. doi:10.1007/s11747-009-0143-3

Meyer, D., Zeileis, A., & Hornik, K. (2015). vcd: Visualizing categorical data. R package version 1.4-1. Retrieved from https://cran.r-project.org/web/packages/vcd/vcd.pdf

Mohammad, S. M., & Turney, P. D. (2010). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. *Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (pp. 26-34). Los Angeles, California: NAACL HLT 2010.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2006). *Introduction to linear regression analysis* (4th ed.). Hoboken, NJ: John Wiley & Sons.

Morgan, N., Pritchard, A., & Piggott, R. (2002). New Zealand, 100% pure. The creation of a powerful niche destination brand. *The Journal of Brand Management, 9*(4), 335-354. doi:10.1057/palgrave.bm.2540082

Muniz Jr., A. M., & O'Guinn, T. C. (2001). Brand community. *Journal of Consumer Research, 27*(4), 412-432. doi:10.1086/319618

Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2015). *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining* (1st ed.). Chichester: John Wiley & Sons Ltd.

Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of Statistical Analysis and Data Mining Applications* (1st ed.). Burlington: Elsevier Inc.

Oliveira, G. H., & Welch, E. W. (2013). Social media use in local government: Linkage of technology, task, and organizational context. 30(4),. *Government Information Quarterly, 30*(4), 397-405. doi:10.1016/j.giq.2013.05.019

O'Reilly, T. (2007). What is Web 2.0: Design patterns and business models for the next generation of software. *Communications & strategies, 65*(1), 17-37. Retrieved from http://ssrn.com/abstract=1008839

Revelle, W. (2016). psych: Procedures for psychological, psychometric, and personality. R package version 1.6.6. Retrieved from https://cran.r-project.org/web/packages/psych/psych.pdf

Rockwell, R. C. (1975). Assessment of multicollinearity: The Haitovsky test of the determinant. *Sociological Methods & Research, 3*(3), 308-320.

Sabate, F., Berbegal-Mirabent, J., Cañabate, A., & Lebherz, P. R. (2014). Factors influencing popularity of branded content in Facebook fan pages. *European Management Journal, 32*(6), 1001-1011. doi:10.1016/j.emj.2014.05.001

Saldaña, J. (2013). *The Coding Manual for Qualitative Researchers* (2nd ed.). London: SAGE Publications Ltd.

Smith, A. N., Fischer, E., & Yongjian, C. (2012). How does brand-related user-generated content differ across YouTube, Facebook, and Twitter? *Journal of Interactive Marketing, 26*(2), 102-113. doi:10.1016/j.intmar.2012.01.002

Statistics Denmark. (2015). *Statistical Yearbook 2015.* Copenhagen: Statistics Denmark.

Statistics Denmark. (2016). *Denmark in figures 2016.* Copenhagen: Statistics Denmark.

Statistics Estonia. (2015). *Statistical yearbook of Estonia.* Tallinn: Statistics Estonia.

Stieglitz, S., & Dang-Xuan, L. (2013). Social media and political communication: a social media analytics framework. *Social Network Analysis and Mining, 3*(4), 1277-1291. doi:10.1007/s13278-012-0079-3

Street, J. (2011). *Mass media, politics and democracy* (2nd ed.). Basingstoke: Palgrave.

Tooman, H., & Müristaja, H. (2014). Developing Estonia as a Positively Surprising Tourist Destination. In C. Costa, E. Panyik, & D. Buhalis (Eds.), *European Tourism Planning and Organisation Systems: The EU Member States.* (pp. 106-117). Aspects of Tourism 61. Bristol: Channel View Publications.

Turner, J. H., & Stets, J. E. (2005). *The Sociology of Emotions.* Cambridge: Cambridge University Press.

Van den Berg, L., & Braun, E. (1999). Urban competitiveness, marketing and the need for organising capacity. *Urban studies, 36*(5-6), 987-999. doi:10.1080/0042098993312

van den Boogaart, K. G., Tolosana, R., & Bren, M. (2013). compositions: Compositional Data Analysis. R package version 1.40-1. Retrieved from https://cran.r-project.org/web/packages/compositions/compositions.pdf

Vanolo, A. (2008). The image of the creative city: Some reflections on urban branding in Turin. *Cities, 25*(6), 370-382. doi:10.1016/j.cities.2008.08.001

VisitCopenhagen. (2016, May 6). *VisitCopenhagen's editorial policy*. Retrieved from VisitCopenhagen The Official Website: http://www.visitcopenhagen.com/editorialline

VisitTallinn. (2016, May 14). *Meie tegevused*. Retrieved from VisitTallinn: http://www.visittallinn.ee/est/turismiprofessionaalile/meie-tegevused

Wickham, H. (2016). plyr: Tools for splitting, applying and combining data. R package version 1.8.4. Retrieved from https://cran.r-project.org/web/packages/plyr/plyr.pdf

Wickham, H., & Chang, W. (2016). devtools: Tools to make developing R packages easier. R package version 1.11.1. Retrieved from https://cran.r-project.org/web/packages/devtools/devtools.pdf

Wickham, H., & Chang, W. (2016). ggplot2: An implementation of the grammar of graphics. R package version 2.1.0. Retrieved from https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf

Wilson, R. E., Gosling, S. D., & Graham, L. T. (2012). A review of Facebook research in the social sciences. *Perspectives on psychological science, 7*(3), 203-220. doi:10.1177/1745691612442904

Yan, J. (2011). Social media in branding: Fulfilling a need. *The journal of brand management, 18*(9), 688-696. doi:10.1057/bm.2011.19

Ye, Q., Law, R., Gu, B., & Chen, W. (2011). The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human Behavior, 27*(2), 634-639. doi:10.1016/j.chb.2010.04.014

yhat. (2013). Fitting & interpreting linear models in R. Retrieved from http://blog.yhat.com/posts/r-lm-summary.html

Appendices

Appendix A1: Facebook data sample

from_name	Message	created_time	type	likes_count	comments_count	shares_count
VisitCopenhagen	The royal family greets the public from the balconies at their winter home Amalienborg Palace in Copenhagen during Queen Margrethe's jubilee. Photo by Stine Avnbøl.	2012-01-15T11:00:00+0000	photo	0	0	0
VisitCopenhagen	Attention foodies: We picked out some of the best restaurants that opened in Copenhagen this year. See the list here:	2015-11-14T16:19:00+0000	link	61	3	8
VisitCopenhagen	Two foreign design students in Copenhagen created this device that pronounces complicated tongue-twister street names in the city. Have you ever had trouble pronouncing Danish words or street names? And would a device like this be helpful? Watch the video and read the full story at http://www.cphpost.dk/news/local/city%E2%80%99s-talking-signs-take-internet-storm http://vimeo.com/45747333	2012-07-21T19:49:36+0000	video	84	10	26
Tallinn	AROUND TOOMPEA Look With New Eyes travel blog presents a beautiful photo essay about Tallinn. Rebecca, the author of the story has discovered exactly the same places in Tallinn Old Town that locals would recommend for your first visit. Thank you, Rebecca, for tha story! Read more bit.ly/1E46qBB Photos by Rebecca/Look withnew eyes.	2015-03-08T20:12:17+0000	photo	162	2	12
Tallinn	HAPPY INDEPENDENCE DAY, ESTONIA! Today we celebrate the 97th anniversary of the Republic of Estonia. The day started with the traditional ceremony of raising the blue, black, and white flag of Estonia on the top of Tall Hermann and continues with festive activities all over the country and in every family. How do you celebrate the birthday of Estonia? Please share your photos and stories.	2015-02-24T09:43:51+0000	photo	847	15	234

Appendix A2: Manual analysis data sample

Message	topic	code	sign	caps	tone	sent	intensity	sec
WHAT DO YOU THINK OF TALLINN? Have you noticed that we have updated our Facebook look to bring you the latest on what's happening in this vibrant capital city? We hope you like it, but tell us your thoughts :)	tourism	tourism	:)	1	1	pride	2	
AND THE WINNERS ARE... Elisabeth Sinipalu, you won 2 tickets to Tallinn Star Weekend 27.06! Terje Kurikoff, you are the lucky winner of 2 tickets to the concert of Andrea Bocelli! Congratulations! You will receive your tickets on e-mail before 12pm 27.06.	competition	culture	…; !	1	1	friendly	3	

Hi, Heiki Järveveer! As a 4000th VisitTallinn' friend, you and three of your friends can enjoy the benefits of 24 hour VIP Tallinn Cards. Please send us your contact details to tourism.marketing@tallinnlv.ee to arrange the hand over of the cards. Congratulations on behalf of Tallinn City Tourist Office and Convention Bureau!	competition	tourism	!	0	1	friendly	3
is happening tonight! Live jazz in city centre and summer parties everywhere - Sankt Hans Torv in Nørrebro and Nyhavn just to mention a few :)	event	culture	!; :)	0	1	anticipation	3
is hoping everyone had a happy Easter :)	holidays	holidays	:)	0	1	friendly	2

Appendix A3: Linear regression data sample

target	type	ct	length	month	hour	year	weekday
0.641033	photo	1	54	8	8	2015	Monday
0.247971	photo	1	19	1	19	2016	Thursday
0.930103	link	0	10	7	10	2010	Monday
0.738871	link	0	21	5	8	2011	Wednesday
0.852931	photo	0	16	5	12	2011	Sunday

Appendix B1: Multiple linear regression results

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.92780 -0.10744  0.04565  0.14952  0.53644

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.013e+00  9.249e-02  10.955  < 2e-16 ***
typelink         -1.228e-01  5.919e-02  -2.075 0.038018 *
typephoto        -2.336e-01  5.898e-02  -3.960 7.62e-05 ***
typestatus       -1.516e-01  6.144e-02  -2.467 0.013670 *
typevideo        -1.413e-01  6.044e-02  -2.337 0.019482 *
ct1               2.130e-02  7.559e-03   2.818 0.004854 **
length            1.468e-03  1.835e-04   8.001 1.60e-15 ***
month02           1.792e-05  1.681e-02   0.001 0.999150
month03          -2.538e-02  1.604e-02  -1.582 0.113634
month04          -4.718e-02  1.658e-02  -2.845 0.004459 **
month05          -1.911e-02  1.704e-02  -1.122 0.262005
month06          -1.197e-02  1.799e-02  -0.666 0.505638
month07          -1.865e-02  1.754e-02  -1.063 0.287821
month08          -2.746e-02  1.682e-02  -1.633 0.102568
month09          -4.028e-02  1.705e-02  -2.363 0.018190 *
month10          -3.272e-02  1.693e-02  -1.933 0.053364 .
month11          -4.230e-02  1.693e-02  -2.500 0.012476 *
month12          -2.614e-02  1.716e-02  -1.523 0.127787
hour01           -1.081e-01  1.052e-01  -1.028 0.303891
hour02            3.268e-02  1.175e-01   0.278 0.780997
hour03           -2.491e-01  1.106e-01  -2.253 0.024317 *
hour04           -1.101e-01  9.802e-02  -1.123 0.261327
hour05           -2.902e-02  7.010e-02  -0.414 0.678930
```

```
hour06            1.737e-02  6.738e-02   0.258 0.796528
hour07            1.769e-02  6.721e-02   0.263 0.792426
hour08            2.474e-02  6.696e-02   0.370 0.711749
hour09            4.039e-02  6.728e-02   0.600 0.548350
hour10            3.231e-02  6.760e-02   0.478 0.632689
hour11            4.769e-03  6.748e-02   0.071 0.943663
hour12            3.480e-03  6.723e-02   0.052 0.958722
hour13            3.445e-02  6.736e-02   0.511 0.609061
hour14            6.344e-03  6.699e-02   0.095 0.924557
hour15           -4.162e-02  6.703e-02  -0.621 0.534647
hour16           -5.456e-03  6.728e-02  -0.081 0.935369
hour17           -2.795e-02  6.704e-02  -0.417 0.676774
hour18           -2.276e-02  6.714e-02  -0.339 0.734663
hour19           -2.389e-02  6.772e-02  -0.353 0.724258
hour20            2.481e-02  6.857e-02   0.362 0.717503
hour21            4.766e-02  7.188e-02   0.663 0.507319
hour22           -3.201e-02  7.439e-02  -0.430 0.667036
hour23           -2.390e-02  8.201e-02  -0.291 0.770692
year2010         -2.207e-02  2.583e-02  -0.854 0.393037
year2011         -2.838e-02  2.581e-02  -1.100 0.271516
year2012         -4.427e-02  2.529e-02  -1.750 0.080147 .
year2013         -8.903e-02  2.598e-02  -3.427 0.000616 ***
year2014         -2.141e-01  2.634e-02  -8.130 5.66e-16 ***
year2015         -2.177e-01  2.607e-02  -8.351  < 2e-16 ***
year2016         -3.511e-01  2.869e-02 -12.239  < 2e-16 ***
weekdayMonday     3.254e-03  1.204e-02   0.270 0.786970
weekdaySaturday   1.010e-02  1.350e-02   0.748 0.454346
weekdaySunday    -1.443e-02  1.333e-02  -1.083 0.278878
weekdayThursday   6.639e-03  1.201e-02   0.553 0.580306
weekdayTuesday   -1.076e-02  1.222e-02  -0.881 0.378575
weekdayWednesday  3.019e-02  1.191e-02   2.534 0.011306 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2169 on 4029 degrees of freedom
Multiple R-squared:  0.2641,    Adjusted R-squared:  0.2545
F-statistic: 27.29 on 53 and 4029 DF,  p-value: < 0.001
```

Appendix B2: Simple linear regressions results

```
#### Predictor: City ####
Residuals:
     Min       1Q   Median       3Q      Max
-0.74389 -0.13915  0.08411  0.20032  0.29588

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.743890   0.005039 147.614  < 2e-16 ***
ct1         -0.039766   0.008020  -4.958  7.4e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2505 on 4081 degrees of freedom
Multiple R-squared:  0.005988,  Adjusted R-squared:  0.005744
F-statistic: 24.58 on 1 and 4081 DF,  p-value: 7.404e-07



#### Predictor: Type ####
Residuals:
     Min       1Q   Median       3Q      Max
```

```
-0.79268 -0.12824  0.07701  0.17763  0.32363

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.97451    0.06481  15.035  < 2e-16 ***
typelink    -0.17364    0.06520  -2.663  0.00777 **
typephoto   -0.29821    0.06500  -4.588 4.61e-06 ***
typestatus  -0.11359    0.06708  -1.693  0.09047 .
typevideo   -0.19338    0.06665  -2.901  0.00373 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2425 on 4078 degrees of freedom
Multiple R-squared:  0.0691,    Adjusted R-squared:  0.06818
F-statistic: 75.67 on 4 and 4078 DF,  p-value: < 0.001
```

Predictor: Length

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.88230 -0.14172  0.08234  0.20140  0.29373

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.7016292  0.0076210  92.065  < 2e-16 ***
length      0.0007728  0.0001901   4.066 4.88e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2508 on 4081 degrees of freedom
Multiple R-squared:  0.004034,  Adjusted R-squared:  0.00379
F-statistic: 16.53 on 1 and 4081 DF,  p-value: 4.879e-05
```

Predictor: Year

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.81093 -0.11246  0.06072  0.15247  0.48368

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.880709   0.021785  40.427  < 2e-16 ***
year2010    -0.008025   0.025477  -0.315  0.75278
year2011    -0.012882   0.024457  -0.527  0.59842
year2012    -0.063864   0.023280  -2.743  0.00611 **
year2013    -0.136743   0.023230  -5.887 4.26e-09 ***
year2014    -0.224830   0.023477  -9.576  < 2e-16 ***
year2015    -0.243378   0.023358 -10.419  < 2e-16 ***
year2016    -0.379713   0.024804 -15.308  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2253 on 4075 degrees of freedom
Multiple R-squared:  0.1968,    Adjusted R-squared:  0.1954
F-statistic: 142.6 on 7 and 4075 DF,  p-value: < 0.001
```

Predictor: Month

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.75007 -0.13946  0.08252  0.19619  0.33155
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.711086   0.013323  53.374  < 2e-16 ***
month02     -0.001395   0.019199  -0.073  0.94207
month03     -0.036124   0.018298  -1.974  0.04843 *
month04     -0.043102   0.018828  -2.289  0.02211 *
month05      0.035740   0.019065   1.875  0.06091 .
month06      0.038280   0.020226   1.893  0.05848 .
month07      0.043919   0.019733   2.226  0.02609 *
month08      0.054251   0.018635   2.911  0.00362 **
month09      0.028703   0.018937   1.516  0.12966
month10      0.032761   0.018749   1.747  0.08064 .
month11      0.019889   0.018868   1.054  0.29190
month12      0.047168   0.018965   2.487  0.01292 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2496 on 4071 degrees of freedom
Multiple R-squared:  0.01558,   Adjusted R-squared:  0.01292
F-statistic: 5.856 on 11 and 4071 DF,  p-value: 1.677e-09


#### Predictor: Weekday ####
Residuals:
     Min       1Q   Median       3Q      Max
-0.75671 -0.13594  0.08202  0.20061  0.32582

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.736690   0.009491  77.617  < 2e-16 ***
weekdayMonday   -0.002788   0.013793  -0.202    0.840
weekdaySaturday -0.014560   0.015392  -0.946    0.344
weekdaySunday   -0.062574   0.015086  -4.148 3.42e-05 ***
weekdayThursday -0.005098   0.013781  -0.370    0.711
weekdayTuesday  -0.015875   0.013975  -1.136    0.256
weekdayWednesday 0.022104   0.013637   1.621    0.105
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2504 on 4076 degrees of freedom
Multiple R-squared:  0.008042,  Adjusted R-squared:  0.006582
F-statistic: 5.508 on 6 and 4076 DF,  p-value: 1.082e-05



#### Predictor: Hour ####
Residuals:
     Min       1Q   Median       3Q      Max
-0.79343 -0.13713  0.07522  0.18530  0.45032

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.752976   0.074024  10.172   <2e-16 ***
hour01      -0.199201   0.118703  -1.678   0.0934 .
hour02       0.027328   0.132418   0.206   0.8365
hour03      -0.277779   0.124601  -2.229   0.0258 *
hour04      -0.152620   0.110349  -1.383   0.1667
hour05      -0.028471   0.078616  -0.362   0.7173
hour06       0.007112   0.075722   0.094   0.9252
hour07       0.036750   0.075562   0.486   0.6267
hour08       0.034143   0.075310   0.453   0.6503
hour09       0.048635   0.075688   0.643   0.5205
hour10       0.043802   0.076073   0.576   0.5648
```

```
hour11       0.004495   0.075956   0.059   0.9528
hour12       0.012121   0.075688   0.160   0.8728
hour13       0.013561   0.075737   0.179   0.8579
hour14      -0.040690   0.075290  -0.540   0.5889
hour15      -0.104762   0.075263  -1.392   0.1640
hour16      -0.075126   0.075490  -0.995   0.3197
hour17      -0.072881   0.075251  -0.968   0.3329
hour18      -0.086799   0.075343  -1.152   0.2494
hour19      -0.105966   0.075920  -1.396   0.1629
hour20      -0.049640   0.076877  -0.646   0.5185
hour21       0.065006   0.080969   0.803   0.4221
hour22      -0.011577   0.083585  -0.139   0.8899
hour23       0.033396   0.092159   0.362   0.7171
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2455 on 4059 degrees of freedom
Multiple R-squared:  0.05038,   Adjusted R-squared:  0.045
F-statistic: 9.363 on 23 and 4059 DF,  p-value: < 0.001
```

Appendix B3: Simple linear regressions results (sentiment)

```
#### Predictor: Tone - Target: Likes ####
Residuals:
     Min       1Q   Median       3Q      Max
-0.10963 -0.04806 -0.02507  0.00797  0.88462

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.11538    0.01983   5.819 2.15e-08 ***
tone1       -0.04433    0.02155  -2.057   0.0409 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1139 on 213 degrees of freedom
Multiple R-squared:  0.01948,   Adjusted R-squared:  0.01487
F-statistic: 4.231 on 1 and 213 DF,  p-value: 0.04091

#### Predictor: Intensity - Target: Likes ####
Residuals:
     Min       1Q   Median       3Q      Max
-0.08381 -0.04983 -0.02493  0.00260  0.91519

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0848112  0.0194569   4.359 2.04e-05 ***
intensity2  -0.0139018  0.0225493  -0.617    0.538
intensity3  -0.0009989  0.0234188  -0.043    0.966
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1151 on 212 degrees of freedom
Multiple R-squared:  0.003329,  Adjusted R-squared:  -0.006074
F-statistic: 0.3541 on 2 and 212 DF,  p-value: 0.7023

#### Predictor: Tone - Target: Comments ####
Residuals:
     Min       1Q   Median       3Q      Max
-0.16883 -0.05808 -0.02512  0.02390  0.83117
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.16883    0.01818   9.287  < 2e-16 ***
tone1       -0.09976    0.01976  -5.049 9.52e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1044 on 213 degrees of freedom
Multiple R-squared:  0.1069,    Adjusted R-squared:  0.1027
F-statistic: 25.49 on 1 and 213 DF,  p-value: 9.52e-07
```

Predictor: Intensity - Target: Comments

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.13972 -0.06000 -0.02423  0.02792  0.86028

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.13972    0.01824   7.660 6.54e-13 ***
intensity2  -0.06193    0.02114  -2.930  0.00376 **
intensity3  -0.07153    0.02195  -3.258  0.00131 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1079 on 212 degrees of freedom
Multiple R-squared:  0.05078,   Adjusted R-squared:  0.04182
F-statistic: 5.671 on 2 and 212 DF,  p-value: 0.00399
```

Predictor: Tone - Target: Shares

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.04978 -0.02865 -0.02865 -0.02865  0.97135

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.04978    0.02233   2.230   0.0268 *
tone1       -0.02113    0.02427  -0.871   0.3848
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1283 on 213 degrees of freedom
Multiple R-squared:  0.003549,  Adjusted R-squared:  -0.00113
F-statistic: 0.7585 on 1 and 213 DF,  p-value: 0.3848
```

Predictor: Intensity - Target: Shares

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.04670 -0.04670 -0.02521 -0.01837  0.97479

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.018367   0.021681   0.847    0.398
intensity2  0.006843   0.025127   0.272    0.786
intensity3  0.028336   0.026096   1.086    0.279

Residual standard error: 0.1283 on 212 degrees of freedom
Multiple R-squared:  0.007983,  Adjusted R-squared:  -0.001376
F-statistic: 0.853 on 2 and 212 DF,  p-value: 0.4276
```

Appendix B4: Multiple linear regression results (per city)

```
### City: Copenhagen ####
Residuals:
     Min       1Q   Median       3Q      Max
-0.80444 -0.08948  0.03924  0.12870  0.52737

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.1110471  0.1655454   6.711 2.40e-11 ***
typelink         -0.1032193  0.1449625  -0.712  0.47651
typephoto        -0.2229948  0.1449374  -1.539  0.12404
typestatus       -0.1387741  0.1463313  -0.948  0.34304
typevideo        -0.1194137  0.1455859  -0.820  0.41217
length            0.0011020  0.0002514   4.382 1.22e-05 ***
month02           0.0080185  0.0205716   0.390  0.69673
month03          -0.0323130  0.0197710  -1.634  0.10231
month04          -0.0436464  0.0202989  -2.150  0.03164 *
month05          -0.0216344  0.0202412  -1.069  0.28526
month06          -0.0076402  0.0211474  -0.361  0.71792
month07          -0.0167309  0.0209959  -0.797  0.42561
month08          -0.0200300  0.0208236  -0.962  0.33620
month09          -0.0360543  0.0209133  -1.724  0.08484 .
month10          -0.0331108  0.0207185  -1.598  0.11015
month11          -0.0147478  0.0206018  -0.716  0.47415
month12           0.0020043  0.0210660   0.095  0.92421
hour01           -0.0486563  0.1382607  -0.352  0.72493
hour02           -0.0517694  0.1615742  -0.320  0.74869
hour03           -0.1275952  0.1613113  -0.791  0.42903
hour05           -0.0783204  0.0778922  -1.005  0.31476
hour06           -0.0437324  0.0742547  -0.589  0.55595
hour07           -0.0505570  0.0741344  -0.682  0.49533
hour08           -0.0717846  0.0739215  -0.971  0.33160
hour09           -0.0255907  0.0745646  -0.343  0.73148
hour10           -0.0520052  0.0751394  -0.692  0.48893
hour11           -0.0506010  0.0746986  -0.677  0.49822
hour12           -0.0526010  0.0746546  -0.705  0.48113
hour13           -0.0179065  0.0751514  -0.238  0.81169
hour14           -0.0660195  0.0739031  -0.893  0.37177
hour15           -0.1056726  0.0738027  -1.432  0.15232
hour16           -0.0543408  0.0740544  -0.734  0.46314
hour17           -0.0852905  0.0738681  -1.155  0.24836
hour18           -0.0861715  0.0744247  -1.158  0.24705
hour19           -0.0850172  0.0755856  -1.125  0.26079
hour20           -0.0612800  0.0761928  -0.804  0.42132
hour21           -0.0100898  0.0788471  -0.128  0.89819
hour22           -0.0566066  0.0811214  -0.698  0.48537
hour23           -0.0912081  0.0864176  -1.055  0.29133
year2010         -0.0459000  0.0294442  -1.559  0.11916
year2011         -0.0441972  0.0306625  -1.441  0.14960
year2012         -0.0502099  0.0304876  -1.647  0.09971 .
year2013         -0.0983321  0.0317329  -3.099  0.00197 **
year2014         -0.2806408  0.0323486  -8.676  < 2e-16 ***
year2015         -0.2999882  0.0324064  -9.257  < 2e-16 ***
year2016         -0.4006758  0.0368961 -10.860  < 2e-16 ***
weekdayMonday    -0.0131743  0.0147227  -0.895  0.37097
weekdaySaturday  -0.0275484  0.0163981  -1.680  0.09309 .
weekdaySunday    -0.0257632  0.0164036  -1.571  0.11641
weekdayThursday  -0.0043097  0.0146077  -0.295  0.76799
weekdayTuesday   -0.0346690  0.0150685  -2.301  0.02149 *
weekdayWednesday  0.0195789  0.0144607   1.354  0.17588
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.2031 on 2397 degrees of freedom
Multiple R-squared:  0.3506,    Adjusted R-squared:  0.3368
F-statistic: 25.38 on 51 and 2397 DF,  p-value: < 2.2e-16
```

City: Tallinn

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.72654 -0.11263  0.04768  0.15423  0.52543

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.7984392  0.1506247   5.301 1.32e-07 ***
typelink         -0.1699260  0.0603583  -2.815 0.004934 **
typephoto        -0.2788463  0.0596742  -4.673 3.22e-06 ***
typestatus       -0.1823458  0.0679929  -2.682 0.007398 **
typevideo        -0.1949211  0.0644301  -3.025 0.002524 **
length            0.0010506  0.0002806   3.744 0.000187 ***
month02          -0.0096466  0.0266161  -0.362 0.717075
month03           0.0037447  0.0253303   0.148 0.882491
month04          -0.0287796  0.0264049  -1.090 0.275907
month05           0.0107500  0.0287767   0.374 0.708777
month06          -0.0128436  0.0308576  -0.416 0.677305
month07           0.0071682  0.0292112   0.245 0.806185
month08          -0.0036857  0.0271896  -0.136 0.892189
month09          -0.0166244  0.0278470  -0.597 0.550600
month10          -0.0290716  0.0275060  -1.057 0.290710
month11          -0.0678874  0.0278412  -2.438 0.014863 *
month12          -0.0670013  0.0276090  -2.427 0.015344 *
hour01            0.1602248  0.1805121   0.888 0.374884
hour02            0.3067988  0.1825258   1.681 0.092989 .
hour03           -0.0901139  0.1693115  -0.532 0.594637
hour04            0.0945858  0.1483657   0.638 0.523880
hour05            0.1422591  0.1354941   1.050 0.293911
hour06            0.1737405  0.1320525   1.316 0.188468
hour07            0.2282915  0.1314209   1.737 0.082564 .
hour08            0.2505648  0.1307721   1.916 0.055540 .
hour09            0.2325266  0.1308568   1.777 0.075768 .
hour10            0.2538736  0.1313694   1.933 0.053474 .
hour11            0.2158154  0.1314255   1.642 0.100766
hour12            0.2044272  0.1307052   1.564 0.118010
hour13            0.2417777  0.1307882   1.849 0.064699 .
hour14            0.2433122  0.1315213   1.850 0.064502 .
hour15            0.2135468  0.1316903   1.622 0.105092
hour16            0.1821643  0.1320928   1.379 0.168070
hour17            0.1930196  0.1314714   1.468 0.142263
hour18            0.1884660  0.1311281   1.437 0.150840
hour19            0.2065842  0.1314338   1.572 0.116203
hour20            0.2577635  0.1331365   1.936 0.053035 .
hour21            0.2085851  0.1419547   1.469 0.141929
hour22            0.0386882  0.1497079   0.258 0.796113
hour23           -0.0528459  0.2577698  -0.205 0.837589
year2010          0.0401515  0.0502537   0.799 0.424425
year2011         -0.0009370  0.0483791  -0.019 0.984549
year2012         -0.0206803  0.0453943  -0.456 0.648762
year2013         -0.0495617  0.0456848  -1.085 0.278148
year2014         -0.0856165  0.0456485  -1.876 0.060900 .
year2015         -0.0878945  0.0451899  -1.945 0.051952 .
year2016         -0.2718999  0.0473036  -5.748 1.08e-08 ***
weekdayMonday     0.0195922  0.0193079   1.015 0.310393
weekdaySaturday   0.0264838  0.0219196   1.208 0.227141
weekdaySunday    -0.0138185  0.0212875  -0.649 0.516344
weekdayThursday   0.0214502  0.0194068   1.105 0.269200
weekdayTuesday   -0.0019907  0.0191948  -0.104 0.917412
weekdayWednesday  0.0306078  0.0193384   1.583 0.113679
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2198 on 1582 degrees of freedom
Multiple R-squared:  0.2103,    Adjusted R-squared:  0.1843
F-statistic:   8.1 on 52 and 1582 DF,  p-value: < 2.2e-16
```

Appendix C: R code

```
1  library(stargazer) # Load library to save results to HTML format (used often)
2  library(plyr) # Load library to count frequencies (used often)
3  setwd("C:/Users/C/Desktop/Thesis/") # Set working directory
4
5  #### Download data ####
6  library(devtools) # Load library to use devtools
7  library(Rfacebook) # Load library to download Facebook data
8  library(httr) # Load library to make Rfacebook usable
9  install_github("Rfacebook", "pablobarbera", subdir = "Rfacebook")
10 fb_oauth <- fbOAuth(app_id = ***, # Removed for privacy reasons
11         app_secret = ***, extended_permission = TRUE)
12 save(fb_oauth, file = "fb_oauth") # Save authorization once
13 load("fb_oauth") # Load authorization
14
15 dates <- seq(as.Date("2009/08/01"), as.Date("2016/04/30"), by = "month")
16
17 downloadCop <- list()
18 for (i in 1:length(dates) - 1) {
19   cat(as.character(dates[i]), " ")
20   try(downloadCop[[i]] <- getPage("VisitCopenhagen", fb_oauth, n = 500, since = dates[i], until = date
21 s[i + 1]))
22   cat("\n")
23 }
24
25 downloadTal <- list()
26 for (i in 1:length(dates) - 1) {
27   cat(as.character(dates[i]), " ")
28   try(downloadTal[[i]] <- getPage("VisitTallinn", fb_oauth, n = 500, since = dates[i], until = dates[i
29 + 1]))
30   cat("\n")
31 }
32
33 cop <- do.call(rbind, downloadCop)
34 tal <- do.call(rbind, downloadTal)
35
36 # Now run line 17-31 again with 2009/08/15 until 2016/04/15 to capture February
37 cop <- rbind(cop, do.call(rbind, downloadCop)) # Save February posts Copenhagen
38 tal <- rbind(tal, do.call(rbind, downloadTal)) # Save February posts Tallinn
39
40 cop$ct <- 0; tal$ct <- 1 # Assign city codes
41 all <- rbind(cop, tal) # Bind all posts
42
43 sampleAll <- all[sample(nrow(all), 5), ] # Get random original posts (for Appendix A1)
44 sampleAll <- sampleAll[ ,-c(1, 6, 7, 11, 15)] # Remove unnecessary columns
45 write.csv(sampleAll, "sampleall.csv") # Save data to csv
46
47 all <- all[-which(is.na(all$message)), ] # Remove posts without messages (83)
48
49 #### Clean text ####
50 library(tm) # Load library for tm_map
51 all$message <- iconv(all$message, "latin1", "ASCII", sub = "") # Remove ASCII characters
52 corpus <- VCorpus(VectorSource(all$message)) # Convert to corpus
53 rml <- function(x) gsub("http[^[:blank:]]+", "", x) # Function to remove hyperlinks
54 corpus <- tm_map(corpus, content_transformer(rml)) # Remove hyperlinks
55 corpus <- tm_map(corpus, removePunctuation) # Remove punctuation
56 corpus <- tm_map(corpus, removeNumbers) # Remove numbers
57 corpus <- tm_map(corpus, stripWhitespace) # Remove white space
```

```r
for (i in 1:nrow(all)) all$clean[i] <- corpus[[i]]$content # Add cleaned message to set
for (i in 1:nrow(all)) all$length[i] <- length(unlist(strsplit(all$clean[i], " ")))

#### Add time and length / remove selected posts ####
all$hour <- substr(all$created_time, 12, 13) # Add hour of the day
all$weekday <- weekdays(strptime(all$created_time, "%Y-%m-%dT%H:%M")) # Add day of week
all$month <- substr(all$created_time, 6, 7) # Add month
all$year <- substr(all$created_time, 1, 4) # Add year

manual <- all[which(all$type == "status"), ] # Get status posts for manual analysis
manual <- manual[-which(grepl("http", manual$message)), ] # Remove status with links
write.csv(manual, "manual.csv") # Save as csv file for manual analysis

all <- all[-which(grepl("http", all$message[which(all$type == "status")])), ] # Remove statuses with l
inks (47)
count(all$type) # Count post types
all <- all[-which(all$type == "note"), ] # Remove 'note' (1)

#### Select columns for analysis ####
analysis <- all[ , c("type", "likes_count", "comments_count",
                     "shares_count", "ct", "length", "month", "hour", "year", "weekday")]
for (i in c(1, 5, 7, 8, 9, 10)) analysis[ ,i] <- as.factor(analysis[ ,i]) # To factors
analysisPlot <- analysis # Copy dataset to create plots

#### PCA ####
library(psych) # Load library for KMO()
KMO(analysis[ ,c(2, 3, 4)]) # KMO adequacy
cor(analysis[ ,c(2, 3, 4)]) # Correlations
bartlett.test(analysis[ ,c(2, 3, 4)]) # Bartlett's test
PCA <- princomp(analysis[ ,c(2, 3, 4)], cor = TRUE, scale = TRUE) # Run PCA
summary(PCA) # Show results
PCA$sdev^2 # Eigen value
analysis <- analysis[ ,-c(2, 3, 4)] # Remove individual targets

#### Cramer's V ####
# Run line 87-90 before running Cramer's V code, and run and line 101-111 after Cramer's V code
library(vcd) # Load library to get Cramer's V
cramer <- matrix(ncol = 7, nrow = 7) # Create empty matrix
colnames(cramer) <- rownames(cramer) <- colnames(analysis) # Column and row names
for (j in 1:7) for (i in 1:7) cramer[i, j] <- round(assocstats(table(analysis[ , c(j, i)]))$cramer, 2)

#### Regression models ####
# Run line 77-98 before running regression model code
analysis <- cbind(target = PCA$scores[ ,1], analysis) # Add target factor to dataset

analysis <- analysis[-which(analysis$target < quantile(analysis$target, .105) |
    analysis$target > quantile(analysis$target, .895)), ] # Remove outliers

normalize <- function(x) (x - min(x)) / (max(x) - min(x)) # Function to normalize to 0-1
analysis$target <- normalize(analysis$target) # Normalize target

m <- target ~ . # To run model with all predictors
m <- target ~ type # To run model with single predictor (replace 'type' by predictor)
model <- lm(m, data = analysis) # Run linear regression model
summary(model) # Show results of model

stargazer(anova(model), summary = FALSE, type = "html", out = "anova.html") # Save Anova

#### Assumptions ####
# Run line 101-111 before running assumptions code
standardized.residual <- rstandard(model)
hist(standardized.residual, prob = TRUE) # Show histogram standardized residuals
curve(dnorm(x, mean = mean(standardized.residual), # Add normal curve
            sd = sd(standardized.residual)), add = TRUE)

plot(model) # Show residuals plots
library(car) # Load library for durbinWatsonTest()
durbinWatsonTest(model) # Run Durbin-Watson test
```

```r

VIF <- vif(model) # Get GVIF values
VIF[,3] <- 1 / VIF[,1] # Add 1 / GVIF
stargazer(VIF, summary = FALSE, type = "html", out = "VIF.html") # Save table

sampleAnalysis <- analysis[sample(nrow(analysis), 5), ] # Get random analysis posts (for Appendix A3)
write.csv(sampleAnalysis, "sampleanalysis.csv") # Save data to csv

stargazer(analysis, type = "html", out = "summary_stats.html") # Get summary statistics
library(ggplot2) # Load library to create plots

#### Plot: Frequencies of likes, comments and shares ####
# Run line 80 before running plot code
freqTargets <- rbind(cbind(count(analysisPlot$likes_count), var = "likes"),
  cbind(count(analysisPlot$comments_count), var = "comments"),
  cbind(count(analysisPlot$shares_count), var = "shares"))

lineCont <- function(s, col, xt, xb, xl, yb, yl, legt, leglab) {
  ggplot(s, aes(x = x, y = freq, colour = col)) + geom_line() +
    scale_x_continuous(xt, breaks = xb, lim = xl) + theme_grey(base_size = 22) +
    scale_y_continuous("Frequency", breaks = yb, lim = yl) +
    scale_colour_discrete(name = legt, labels = leglab)
}
lineCont(freqTargets, freqTargets$var, "Quantity per post", seq(0, 500, 125), c(0, 500),
  seq(0, 100, 25), c(0, 100), "Measure",c("Likes", "Comments", "Shares"))

#### Plot: Frequency of length ####
lineCont(count(analysisPlot$length), factor(1), "Post length", seq(0, 120, 30), c(0, 120),
  seq(0, 150, 50), c(0, 150), "", "") + theme(legend.position = "none")

#### Plots: Distributions per hour, weekday, month and year ####
distCat <- function(s, col, xt, yb, yl) {
  ggplot(count(s), aes(x = x, y = freq)) +
    geom_bar(stat = "identity", fill = col) +
    scale_y_continuous(breaks = yb, lim = yl) + labs(x = xt, y = "Frequency") +
    geom_text(aes(label = freq), vjust = -.25, size = 6) + theme_grey(base_size = 22)
}

distCat(analysisPlot$hour, "#00D78C", "Hour of the day", seq(0, 450, 90), c(0, 450))

freqWeekdays <- count(analysisPlot$weekday)
levels(freqWeekdays$x) <- c("Monday", "Tuesday", "Wednesday", "Thursday",
                            "Friday", "Saturday", "Sunday")

distCat(freqWeekdays, "#D73600", "Weekday", seq(0, 1000, 250), c(0, 1000))
distCat(analysisPlot$month, "#65D700", "Month", seq(0, 600, 100), c(0, 600))
distCat(analysisPlot$year, "#D79A00", "Year", seq(0, 1200, 300), c(0, 1200))

#### Plot: Frequency of post types ####
frequenciesType <- count(analysis$type) # Count frequencies
frequenciesType <- frequenciesType[order(-frequenciesType[,2]),] # Sort by frequency

ggplot(frequenciesType, aes(x = 1, y = freq, fill = sort(factor(x)))) +
  geom_bar(width = 1, stat = "identity") + theme_minimal() + coord_polar(theta = "y") +
  theme(axis.title = element_blank(), axis.text.y = element_blank(),
      axis.text.x = element_blank(),
      legend.text = element_text(size = 15), legend.title = element_text(size = 20)) +
  scale_fill_hue(paste("Type of post"),
    labels = paste(frequenciesType$x, " (", frequenciesType$freq, ")", sep = ""))

#### Plot: Normalized dependent variable ####
# Run line 101-107 before creating this plot
ggplot(count(round(analysis$target, 1)), aes(x = x, y = freq, fill = freq)) +
  geom_bar(stat = "identity") +
  scale_x_continuous("Normalized dependent variable",
    breaks = seq(0, 1, 0.1), lim = c(-0.05, 1.05)) +
  scale_y_continuous("Frequency", expand = c(0, 0),
    breaks = seq(0, 1200, 300), lim = c(0, 1200)) +
  theme_grey(base_size = 22) +  scale_fill_gradient(guide = FALSE) +
```

```r
  geom_text(aes(label = freq), vjust = -.25, size = 6)

#### Manual analysis ####
coded <- read.csv("manual.csv", stringsAsFactor = FALSE) # Read coded posts

length(which(coded$sec == "")) # Agreement
mean(coded$tone) # Mean tone
count(coded$tone) # Tone distribution
mean(coded$intensity) # Mean intensity
count(coded$intensity) # Intensity distriubtion
length(which(grepl(":", coded$sign) == TRUE)) # Number of posts with smiley faces
length(which(grepl("!", coded$sign) == TRUE)) # Number of posts with !
count(coded$caps) # Distribution of words in all capital letters (caps)

# For the plot below, create the distCat function (line 157-162) first
distCat(coded$sent, "#2D3AD6", "Emotion", seq(0, 80, 20), c(0, 80)) # Sentiment distribution
write.csv(count(coded$code),"codes.csv") # Save code distribution

#### Expected values for individual predictors ####
mean(analysis$target) # Get overall average of target
meanCity <- ddply(analysis, "ct", summarize, Mean = mean(target)) # Add mean for value
meanCity$OverallMean <- mean(analysis$target) # Add overall mean
meanCity$Difference <- meanCity$Mean - mean(analysis$target) # Add difference
stargazer(meanCity, summary = FALSE, rownames = FALSE,
  type = "html", out = "meanCity.html") # Save table
# For other predictors, replace 'ct' by name of predictor (e.g. 'type'/'year')

#### Regression models sentiment ####
# Run line 198 before running sentiment regression model code
coded[ ,7] <- as.factor(coded[ ,7]) # Change tone to factor
coded[ ,9] <- as.factor(coded[ ,9]) # Change intensity to factor

coded$likes_count <- normalize(coded$likes_count) # Normalize likes
coded$comments_count <- normalize(coded$comments_count) # Normalize comments
coded$shares_count <- normalize(coded$shares_count) # Normalize shares

m <- likes_count ~ tone # Define model (replace likes and tone accordingly: target ~ predictor)
model <- lm(m, data = coded) # Run linear regression model
summary(model) # Show results of model

#### Regression models comparing cities ####
# Run line 238-259 with choosing either Copenhagen (line 240) or Tallinn (line 241), then run line
238-259 again, with choosing the other city
analysis <- all[ , c("type", "likes_count", "comments_count",
  "shares_count", "ct", "length", "month", "hour", "year", "weekday")]
analysis <- analysis[which(analysis$ct == 0), ] # City = Copenhagen
analysis <- analysis[which(analysis$ct == 1), ] # City = Tallinn
for (i in c(1, 5, 7, 8, 9, 10)) analysis[ ,i] <- as.factor(analysis[ ,i]) # To factors

KMO(analysis[ ,c(2, 3, 4)]) # KMO adequacy
cor(analysis[ ,c(2, 3, 4)]) # Correlations
bartlett.test(analysis[ ,c(2, 3, 4)]) # Bartlett's test
PCA <- princomp(analysis[ ,c(2, 3, 4)], cor = TRUE, scale = TRUE) # Run PCA
summary(PCA) # Show results
PCA$sdev^2 # Eigen value
analysis <- analysis[ ,-c(2, 3, 4, 5)] # Remove individual targets and city

analysis <- cbind(target = PCA$scores[ ,1], analysis) # Add target factor to dataset
analysis <- analysis[-which(analysis$target < quantile(analysis$target, .105) |
  analysis$target > quantile(analysis$target, .895)), ] # Remove outliers
analysis$target <- normalize(analysis$target) # Normalize target

m <- target ~ . # Define model
model <- lm(m, data = analysis) # Run linear regression model
summary(model) # Show results of model
```