**Introduction/Motivation**

Reviews play an important role in determining how users feel about a certain product. Companies have a wide variety services that are offered and may attract a wide variety of consumers. Social media is a product that is readily available to any individual that has access to the internet. Among all the popular social media websites, Facebook is the largest. As of October 2020, Facebook reported 2.7 billion monthly active users [1]. Knowing how the customers feel about the different services and policies that are being implemented by Facebook may provide insightful information on how people feel about the company. Segmenting reviews by topics may provide additional benefits by allowing a company to identify key issues and prioritize which of those issues need to be focused on primarily or in some cases to be ignored. Segmenting the content of the reviews can help identify for example how many reviews are about technical issues vs. how many reviews are about lost access to account vs how many reviews are about the user experience on Facebook. In a broader scope, developing techniques for better text analysis can improve recommendation algorithms for a company’s product(s). Another viable concern is that anyone can post a review about a company, product, or service online. Trustpilot is a Danish consumer review website where people can leave reviews about businesses and where nearly 1 million reviews are posted each month [2]. If such review websites are to have credibility in the quality of their reviews, then it is important for the company to detect and remove fake reviews that do not give a genuine insight into the user’s experience and remove reviews that are offensive in nature as they provide no additional benefits to those who read the review. On their business site, Trustpilot explains how the reviews are monitored to make the experience of leaving a review as fair and transparent as possible. They provide some information on how they monitor for fake reviews and those that violate their community guidelines [3]. My first research question is: what are the different topics (technical issues, censoring, etc.) that the Facebook reviews are written on Trustpilot? My second research question is: how these reviews can be explored via text and sentiment analysis to detect reviews that are not genuine?

**Review of Existing Theory and Evidence**

There is an ongoing interest in using sentiment analysis, cluster analysis and topic modelling to evaluate the content that is written in online reviews. According to Fang and Zhan [4], online reviews may contain flaws that can hinder sentiment analysis. People can post whatever they want online without any repercussions. This could be a fake review where the user either never had experience using a product or provides a non-genuine experience review. In addition, spammers leave spam messages in the reviews that have little to no relation to the actual product. In their paper Fang and Zhan mention the concept of ground truth which classifies content into categories such as positive, neutral, and/or negative. They suggest review ratings can be used as the ground truth. I intended to explore if a relationship exists between assigned sentiment ratings and the ratings of the review. Comparing the sentiment of an individual review to the overall rating left by the reviewer may help with classifying what the review is about. This can be used as a filter to detect fake reviews in the dataset. Detecting these outliers is important because it may be possible to detect fake and/or spam reviews in the dataset. According to Fang and Zhan, they found some limitations when dealing with sentiment analysis at the review-level. They found that categorizing reviews becomes difficult when classifying reviews to their specific star-scaled ratings which caused them to report F1 scores less than 0.5. Another limitation that they suggest considering when performing sentiment analysis is that it may not work on reviews that have no tokens that indicate a specific sentiment. To try and account for these limitations in my exploration of data, I use the overall sentiment of a review as variable that can be looked at in correlation with pre-existing variables in the dataset such as rating and the number of reviews that are left by the reviewer on the Trustpilot website overall.

In the paper written by Hiremath and their colleagues [5], they indicate that the structure of the review system can also dictate how these reviews should be analyzed. Some review formats offer features that can be used in cluster analysis. A feature is a list of pros and cons for a specific product in which cluster analysis may be a more appropriate methodology of use rather than text analysis. Features can be classified into different clusters and shown in the paper. Unfortunately, the Trustpilot reviews do not come in this format that accommodates features. Trustpilot reviews are written in a free-form text format where only additional review information that is provided is the rating of the review. Regardless of this limitation it is still possible to perform cluster analysis on the dataset using features such as the date when the review was submitted, the number of reviews on Trustpilot that are associated with a specific account, the rating given by the reviewer, the overall measured sentiment of the review, and classification categories (positive, neutral, or negative) for the overall sentiment for a review. According to Hiremath and their colleagues, using these features can help cluster reviews into categories such as: most significant review, more significant review, significant review, and/or insignificant review. In my project I use cluster analysis results to identify how many topics the reviews may be written on in the Facebook review dataset. I use cluster analysis results more as a guide that tells me how many topics I should expect prior to performing topic modeling techniques and most importantly the data that is used in cluster analysis is numeric in nature and does not contain text (which is what is used in sentiment analysis and topic modeling). One final conclusion made by Hiremath and their colleagues is that there is a significant number of features that they found that belonged to insignificant group of reviews and they recommend that these reviews be excluded for further analysis.

In a Towards Data Science article written by Ng [6], they show how categorizing reviews into topics may be beneficial in not only describing the content that is found within the reviews (descriptive analytics) but also how these topics can be used in predictive analytics and prescriptive analytics. Using review topics in predictive analytics can give insights about a rating that the reviewer is likely to give when writing a review. This is useful because it can be applied by companies like Trustpilot that would want to check if the ratings of the review match the context of the review itself. This can help filter for those reviews that are rated inappropriately and increase confusion for those who read the reviews when researching a product to buy and/or use. Review topics can also help predict whether a future consumer will buy and/or use a product based on the reviews that are written on that product. This can inform the company that offers/makes the product about things like predicted earnings/losses for a product/service that is offered and this can also advise on questions in the realm of inventory management. Using review topics in prescriptive analytics also has its benefits because it has the potential for companies to find consumers that may be more inclined to use their products and/or services.

As a result of the background evidence provided for my data exploration, I will attempt at providing an even-handed presentation of different methodologies that can be applied to segment reviews into separate topic and theme categories.

**Methodology**

**i.Data**

The reviews dataset consists of reviews that are written by Trustpilot members on their experience with Facebook. Trustpilot is a third-party site where these reviews are written by anyone who provides a valid email address or uses their Facebook account. Trustpilot monitors reviews on their site via several methods including software detection programs that detect anomalies in their data and rely on their community members to flag and report reviews that are misleading, spam, and/or use hate speech. The dataset was obtained from Kaggle [7] and the original dataset consists of 3261 reviews. The time period of the reviews ranges from February 2011 to October 2020.

**ii.Variables**

After data cleaning and manipulation, two datasets were generated. The dataset that was used in topic modelling is called “sentiment\_analysis\_reviews” and has the following variables: Time (when the review was written on Trustpilot), Num\_Reviews (total number of reviews on Trustpilot associated with a particular account), Rating (a score given by the Trustpilot reviewer), id (identification number for a review), review (all the written text including both the title and the body of the review), afinn\_sentiment (total sentiment score of a review generated after sentiment analysis), negative (binary variable indicating overall negative sentiment), neutral (binary variable indicating overall neutral sentiment), and positive (binary variable indicating overall positive sentiment). The dataset used in cluster analysis is called “cluster\_analysis” and has the following variables: Time, Num\_Reviews, Rating, afinn\_sentiment, negative, neutral, and positive.

**iii. Modeling approach**

Three methodologies (sentiment analysis, cluster analysis, and topic modeling) that were covered in class were applied when exploring this dataset. In the sentiment analysis approach, the reviews were tokenized into words and given a sentiment score using the AFINN sentiment lexicon which contains a collection of words expressing varying degrees of sentiment). The sentiment score for each word was then added together for each individual review and the total sentiment score was obtained for the review. Using the sentiment scores, I explored the dataset for potential outliers or anomalies that would indicate a fake review. In the process, I decided to filter for reviews where the Num\_Reviews was between 25 to 200 and the afinn\_sentiment -50 and 25. Looking at the original datset with outliers (Figure 1) we can see that there were reviews that had a very positive afinn\_sentiment but with rating of 1 or a review with a very negative afinn\_sentiment but with a rating of 5. For reviews with a rating of 1-2 and afinn\_sentiment of -26 to -6, I categorized these reviews into negative\_sentiment variable. For reviews with a rating of 3 and afinn\_sentiment of -5 to 5, I categorized these reviews into neutral\_sentiment variable. For reviews with a rating 4 to 5 and afinn\_sentiment of 6 to 13, I categorized these reviews into positive\_sentiment variable. This was done to not have a discrepancy between Rating and afinn\_sentiment variables that was discovered during EDA. The new dataset not including the outliers/faulty reviews and their respective Ratings/afinn\_sentiment values can be seen in Figure 2, please refer to the end of this document.

For cluster analysis, the variables Time, Num\_Reviews, Rating, afinn\_sentiment, negative, neutral, and positive were used and standardized to the same scale. K-means clustering using the Elbow, Silhouette, and Gap Statistic methods were used to determine the number of clusters in the dataset.

For topic modeling, the “sentiment\_analysis\_reviews” dataset was used. An LDA with Gibbs sampling was performed for both k = 2 and k = 7 to obtain top words per topic(betas) and top topics per entire dataset (gammas) [8]. A k = 2 was used based on the results obtained in cluster analysis whereas k = 7 was used based on results from perplexity scores obtained using the test data subset of the “sentiment\_analysis\_reviews” dataset.

**Results**

Based on the sentiment analysis scores (afinn\_sentiment variable), it was possible to detect problematic reviews. For example, reviews that had a low rating score but had positive sentiment or reviews that had a high rating score but had low positive sentiment expressed were detected and removed. This proved to be effective because this allowed to detect reviews that were for example written in both English and Danish where positive words/sentiments in the review were written in English and the negative words/sentiments were written in Danish. Why exactly such reviews exist is not completely understood but it could be a tactic used by the reviewer to minimize the risk of their review being taken down due to violating the community guidelines. Additionally, to try and increase the credibility of the reviews, reviews that written by users who had 25 to 200 reviews associated with their Trustpilot account were kept. This was done to minimize the occurrence of poor quality and spam reviews but is no way a perfect mechanism. The result is that out of the original 3261 reviews, 58 were only selected for further analysis.

In the cluster analysis approach, three methods were used: Elbow, Silhouette, and Gap Statistic. Looking at the Elbow method results in Figure 3; it is suggested that the optimal number of clusters is 2. Looking at the Silhouette method results in Figure 4; it is suggested that the optimal number of clusters is also 2. Looking at the Gap Statistic method results in Figure 5; it is suggested that the optimal number of clusters is also 2. Based on these results k = 2 was used in topic modeling.

Perplexity which is measure of how well a probability model fits new data was also used to find optimal k (number of topics) in the topic modeling part of this project. Lower perplexity scores indicate better model fit. Looking at perplexity by the number of topics in Figure 6, 7 to 11 number of topics seem reasonable to explore further. Based on the results obtained from cluster analysis and perplexity per number of topics, I decided to create two LDA models where k = 2 and k = 7, respectively. Due to time constraints, I decided to analyze more in detail the LDA model where k = 2 by looking at the top 30 words per topics 1 and 2 (Figures 7,8). Reading reviews with the top 10 gamma values (top reviews per topic) for both topics 1 and 2, I came to the following conclusions: Topic 1 was about security issues, scammers, and privacy issues while Topic 2 was about censorship of content, poor social experience (too many ads, cannot see friend’s content), poor community standards, poor customer service, and poor experience with Facebook games. Additional graphics for both k =2 and k =7 can be seen in the attached R code, for further exploration.

**Conclusions**

**i.Key takeaways/conclusions**

Exploring data using sentiment analysis as a first step can be beneficial when exploring text data, especially if a variable such as rating is present and can be logically correlated with. Creating sentiment scores per review allowed for the detection of reviews that were illogical and hard to interpret. If these outliers (fake or faulty reviews) are kept in the dataset, they can complicate further analysis and modeling that may be done in the future as in the case provided by [6] where topic modeling is used in predictive and prescriptive model development. Sentiment score also allowed for the creation of new variables that were then used in cluster analysis.

When it comes to cluster analysis, sample size becomes an important assumption to consider. Although there tends to be no exact guideline for the appropriate sample size to be used in cluster analysis, larger sample sizes are often preferred. According to [9], past research suggests sample sizes that are 2^k observations (where k equals the number of variables) are appropriate. Thus, according to these previous findings, the sample size used in my analysis may not be sufficient.

Results obtained from topic modeling are subjective in nature and/or require great expertise. As we saw in the perplexity by number of topics graphic (Figure 6), there can be multiple k values that may seem appropriate. I think a good rule of thumb is that exploring different k values can be beneficial if the problem at hand is of great interest, and that increasing k makes each topic more specific. So when choosing the number of topics, the best resource to rely on and making sound judgement is often based on past experiences and also the context of the problem at hand can be extremely helpful when choosing a number of topics to consider.

**ii.Limitations of current analysis**

One limitation of the analysis is the possibility that random sampling may have not occurred when acquiring this data. Self-selection bias is possible with this kind of data because reviewers with polarizing feelings may be more inclined to leave a review than individuals who may have more neutrally aligned opinions and experiences when it comes to Facebook and the services it offers. In my case, I did not know about Trustpilot until doing this project and would have never left a review unless I felt a strong compelling reason to review Facebook and research for a way to leave a review about Facebook on the web. When it comes to examining the efficiency of upholding the stated Trustpilot community guidelines [3], it is also beneficial to consider removing reviews from analysis that were submitted 2-9 years ago as they may not reflect the performance of the software detection technology that Trustpilot mentions about. In either case, why the reviews that violate community guidelines and come from an older time period remain, is a question that is worth exploring on its own. Although my project was about exploring review dataset and trying to come up with topics for the reviews, it is important to note that my method of filtering out outliers(fake reviews) is by no means ideal and may have unforeseen flaws. One such possible flaw is that by excluding outliers and aggregating further down to reviews that were submitted by accounts with 25 – 200 total reviews, my sample size decreased from 3261 observations to 58 observations. The argument that I would like to make here is that many reviews should not be considered because either they were confusing (due to the contents of the review not corresponding with rating) or offensive in nature and disrupted community guidelines anyways. I think if I had more time, I would definitely experiment with different sample sizes.

**Tables/Figures**

Figure 1 – Original dataset with outliers

Chart, line chart

Description automatically generated

Figure 2 – New dataset without outliers

Chart, line chart

Description automatically generated

**Tables/Figures (continued)**

Figure 3 – Elbow method from cluster analysis

Chart, line chart

Description automatically generated

Figure 4 – Silhouette method from cluster analysis

Chart, line chart

Description automatically generated

**Tables/Figures (continued)**

Figure 5 – Gap Statistic method from cluster analysis

Chart, line chart

Description automatically generated

Figure 6 – Perplexity per number of topics

Chart, scatter chart

Description automatically generated

**Tables/Figures (continued)**

Figure 7 – Top 30 words for topic 1 where k=2

Chart, bar chart

Description automatically generated

Figure 8 – Top 30 words for topic 2 where k=2

Chart, bar chart

Description automatically generated

**References**

[1] Clement, J. “Global Social Media Ranking 2019 | Statistic.” *Statista*, Statista, 2019, www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

[2] “Trustpilot.” *Wikipedia*, 29 Nov. 2020, en.wikipedia.org/wiki/Trustpilot#Criticism. Accessed 10 Dec. 2020.

[3] “Trustpilot - Combating Fake Reviews.” *Trustpilot Business*, www.trustpilot.com/trust/combating-fake-reviews. Accessed 10 Dec. 2020.

[4] Fang, Xing, and Justin Zhan. “Sentiment Analysis Using Product Review Data.” *Journal of Big Data*, vol. 2, no. 1, 16 June 2015, 10.1186/s40537-015-0015-2. Accessed 23 Apr. 2019.

[5] Hiremath, Prakash & Algur, Siddu & Shivashankar, S.. (2010). Cluster Analysis of Customer Reviews Extracted from Web Pages. Journal of Applied Computer Science & Mathematics. 4.

[6] Ng, James. “Natural Language Processing (NLP) to Analyse Product Reviews by Online Shoppers.” *Medium*, 16 Apr. 2020, towardsdatascience.com/natural-language-processing-nlp-analysis-of-product-reviews-by-online-shoppers-7a5966f5e615. Accessed 10 Dec. 2020.

[7] Wolber, Leon. “Facebook\_reviews\_trustpilot.” *Kaggle.com*, www.kaggle.com/leonwolber/facebook-reviews-trustpilot. Accessed 10 Dec. 2020.

**References (continued)**

[8] Tomar, Ankur. “Topic Modeling Using Latent Dirichlet Allocation (LDA) and Gibbs Sampling Explained!” *Medium*, 28 Nov. 2018, medium.com/analytics-vidhya/topic-modeling-using-lda-and-gibbs-sampling-explained-49d49b3d1045. Accessed 10 Dec. 2020.

[9] Idlette-Wilson, Alexis. “Sample Size for Cluster Analysis.” *Medium*, 13 Nov. 2018, medium.com/business-data-quality-analyst/sample-size-for-cluster-analysis-72260a40e41d. Accessed 10 Dec. 2020.

‌

‌

‌

‌

‌