Analysing the Social-Economic Impact of Wireless Mobile Services During and Before COVID-19 Using Topic Modelling and Sentiment Analysis on Tweets

Applied Data Science Masters Thesis

Wehel Hadi
8686807
w.hadi@students.uu.nl

First supervisor
Dr. A.A.A. (Hakim) Qahtan

Second supervisor
Dr. M.W.(Mel) Chekol

Universiteit Utrecht

Department of Information and Computing Sciences
Utrecht University
Netherlands
August 2021

Abstract

Social media platforms can be used as a data source for measuring public opinion on various topics such as wireless mobile services. Twitter is a suitable platform that is able to map the sentiments. In this research the influence of wireless mobile services on values such as user satisfaction (social effect), affordability (economic effect) and willingness (social effect) is researched. This research is conducted through a created system that uses topic modeling and sentiment analysis. HDP, LDA and LSI are the topic models used to map the various topics. While Multinomial Logistic Regression, Naive Bayes, Decision Trees and Random Forest are the sentiment models that map the sentiment per value. All these models are evaluated for their performance with the aim of choosing the best model for the system. This research will also determine the sentiment over time for each value and the sentiment for the companies Mint Mobile and Infinity Mobile. These companies differ from policy, so this analysis provides insight into the influence of company policies on these values. The analysis has shown that the overall sentiment for user satisfaction, affordability and willingness is negative. It can also be seen that the pandemic has played a major role in this negative sentiment. For both willingness and affordability, a clear trend break can be observed at the start of the pandemic. Finally, it is also observable that for all values the company with a more flexible company policy(Mint Mobile) has a less negative sentiment than the company (Infinity Mobile) with a traditional company policy.

Acknowledgements

I would like to thank the people who played an essential role in the completion of this research paper. Without the help of these people, the quality of the report would not have been the same. First of all, I would like to thank my thesis supervisor, Hakim Qahtan, for his guidance and commitment during the thesis project. During the thesis, Mr. Qahtan always took the time to answer all my questions and gave me feedback at all times. I would also like to thank him for joining a new thesis project halfway through the last quarter. This is something that is not taken for granted, and for this I am very grateful. I also want to express my gratitude to the study coordinators, David van Balen, Mel Chekol and Matthieu Brinkhuis, who have put a lot of effort into finding me a new supervisor. I would like to say a special thank you to Mel Chekol for the extra role he has taken as second examiner. Finally, I would like to thank my family for their support during this thesis project. They have always supported me throughout my studies and always believed in my abilities. My accomplishments and success can largely be attributed to them. All in all, I really enjoyed the process of doing research and writing a scientific paper. That's why I look back on it with pleasure.

Contents

Chapter 1

Introduction

Over the past twenty years, wireless communication has undergone a technological revolution. Wireless mobile services have become the fastest growing part of the telecommunications sector[7]. The use of a mobile phone has become an essential part of today's society. Today, 90 percent of the world's population over the age of six has access to a mobile phone, i.e. billions of people[10]. This makes wireless mobile communication a worldwide phenomenon, for developed countries as undeveloped countries. The rise of mobile technology therefore has a direct social-economic impact. Wireless mobile services are thus indispensable in our society today, but this has not always been the case. The wireless technology industry has evolved over the years, leading to a mobile revolution with fundamental and major consequences for the world. However, this revolution has not yet come to an end, innovations are made every year. This requires adaptability from both society and the economy. However, the advantages of these emerging technologies outweighs the disadvantages. A disadvantage of this emerging technology was researched by Subramani Parasuraman. The study of Parasuraman et al. [27] showed that a significant number of the participants of the research had an addiction to mobile phone usage. Another interesting result was that the majority of these participants didn't recognized that they were addicted. On the other hand, the mobile revolution has made it possible to obtain more easily available data so that decision makers can make better decisions. It has also had a positive influence on communication between people, but also between institutions and people. The government had been given a tool that made it easier to reach the people. Companies also had this advantage with regard to their customers[36]. Wireless mobile services, have also made a positive contribution to personal security. Nowadays it is possible to immediately ask for help in dangerous situations with the help of always available internet and mobile phone calls. In addition, the rise of this technology has also created a completely new industry. This has created many jobs for people, which has had a positive impact on employment and the economy[7]. This emerging industry is made up of different companies with different perspectives and ways of providing services. Two companies that differ in their way of delivering services are Infinity mobile and Mint Mobile. While Infinity mobile opts for the traditional fixed plan approach[25], Mint Mobile opts for more flexible plans for their customers[24].

Therefore, this paper examines the social-economic impact of wireless mobile services

for the companies Infinity Mobile and Mint Mobile. These two companies contradict each other in their way of delivering services. This makes it interesting to investigate whether a company's policy influences the sentiment of core values such as user satisfaction (social effect), affordability (economic effect) and willingness (social effect). The research question for this paper is therefore: "How does wireless mobile services impact values as user satisfaction, affordability and willingness?" The analysis will be performed using topic modelling and sentiment analysis on data from twitter. Topic modelling will reflect the various topics discussed in the data. The sentiment analysis will determine what the sentiment is for each of these values. During this analysis, the changes in the sentiment score over time will be determined. Based on that, it will be clear whether COVID-19 affected the sentiment score or not. The first sub question for this paper is therefore:" Did the pandemic change the impact of wireless mobile services on user satisfaction affordability and willingness?" The last sub question of this paper is: "To what extent do company policies influence values as user satisfaction, affordability and willingness for wireless mobile services? "

The remainder of this paper is organized as follows: section 2 is the related work, where the different theories and models of other researchers are discussed. Section 3 is the theoretical background; this chapter will introduce the different models that are used for the analysis. Section 4 is the system architecture, this chapter discusses the different steps of the system; the data collection, data preparation, topic modelling, the sentiment analysis and the evaluation. Section 5 is the evaluation, this chapter describes the evaluation procedure of the system and also discusses the analysis results. Section 6 contains the conclusion. This chapter describes the summary of the research, the limitations of the research, the ethical considerations and the further research that needs to be done. Section 7 is the references and the last section is the appendix. The appendix also contains a link to GitHub for the code.

Chapter 2

Related Work

The aim of this paper is to gain insight into the impact of wireless mobile services on user satisfaction, affordability and willingness (to use). The impact is investigated by using topic models and sentiment analysis techniques. The exact way in which this will be investigated is based on theories, models and concepts. The various studies that contribute to the theoretical framework of this paper will be discussed in this chapter.

User satisfaction, also called customer satisfaction, is one of the values that will be explored in this paper. This value was defined in the study of Lee et al.[20]. According to Lee, user satisfaction is the extent to which the expectations of the product or service are met. If the expectations were higher than the experience, it leads to disappointment. If the service is better than expected, it can lead to a very satisfied customer. Affordability is another value that will be explored. The paper by Pau et al. [30] examined to what extent mobile and media communications services are still affordable for an average family, balancing the necessity of these services and the financial risks. It emerged from this paper that affordability is mainly a balance issue. Affordability is keeping the balance between the important mandatory expenses and the costs for mobile services. The last value that will be examined in this paper is the willingness to use. Hong et al.[14] has researched customer satisfaction and willingness to use self-service kiosks in hotels. In this research, the value of willingness was defined as follows: "An individual's openness to a new opportunity". These values have a direct relationship with society and the economy. Sarwar et al.[33] has investigated the influence of wireless mobile services on society. This research also investigates how this technology will structure the society of the future. This paper has outlined all the pros and cons, however, this study did not compare the societal effects caused by different providers of these services. This is a research gap that will be filled in this paper.

The relevant values should be extracted from the data; this can be achieved using different topic models. Negara et al.[26] analyzed the performance of the Latent Dirichlet Allocation(LDA) method. In this research Indonesian tweets were clustered in four topics; military, technology, economics and sports. These clusters were also visualized. This research showed that LDA was an excellent topic model, the model was found to have a 98 percent accuracy score for the 4 topics. However, the LDA model showed

some problems with non-English tweets. Merchant et al.[23] analyzed the performance of the Latent Semantic Analysis(LSA) method. The aim of this study was to cluster court data in various topics. The clustering of this data was realized with the help of LSA. Clustering this data had major advantages for both lawyers and citizens. This method enabled stakeholders to obtain useful information more quickly. This research showed that this topic modeling method was eventually approved for use by professional lawyers. Hoffman et al.[13] have researched the Hierarchical Dirichlet Process (HDP) method. Hoffmann researched clusters within recorded songs. The aim of his research was to determine similarities between songs using the HDP model and to divide these songs into categories. The result of this research was that the HDP model was a fast and accurate method, which also outperformed the G1 algorithm. This algorithm had a lot of trouble distinguishing the songs that were similar to the other songs.

The already mentioned topic models have yielded good results according to the respective studies. However, one of these models will have to be chosen for this paper. As a result, these models need to be evaluated. In the paper by Stevens et al.[37], various models were compared with the help of an evaluation parameter. This evaluation parameter is called topic coherence. Topic coherence is a measure for the semantic similarity between words in topics. The coherence score varies between 0 and 1, but generally the higher the coherence score, the better the topic model[12]. The average value of this coherence score determines the quality of the relevant topic model. The hyper parameters that ensure that a model functions optimally were also discussed in this paper. The coherence score of a model depends on the number of chosen topics. The optimum number of topics must be found to determine an accurate coherence score. This research compared the LSA, LDA and NMF models. Stevens et al.[37] concluded that each topic model had both advantages and disadvantages, however the LDA model had the highest coherence score followed by the NMF model and then the LSA model. The last step of the topic modeling process is to visualize the results. Jelodar et al.[16] has researched research papers on topic modeling in the period 2003 to 2018, looking at developments and trends around topic modelling. The emphasis of this paper was on the problems that come with the visualization. This paper has included research papers related to different sciences, aiming to cover a spectrum as broad as possible. Jelodar et al.[16] concluded that topic modeling can provide a good view of all terms together, but also the individual documents and their mutual relationships.

The end goal of this paper is to determine the sentiment per value. A sentiment analysis is therefore the next step after the topic modeling. Kaur[17] has researched the sentiment surrounding COVID-19, using a Naive Bayes classifier. Kaur et al.[17] has mined the data from Twitter through the twitter API. This study showed that the accuracy of the neutral sentiment was insufficient. The share of neutral sentiment in the data was too high, thus indicating a malfunction of the Naive Bayes classifier. A research gap in this paper was examining the sentiment over time, which could provide a more insightful view of the sentiment. This research gap will also be filled in this paper. Parmar et al.[29] has researched the sentiment regarding movie reviews. The researchers performed this sentiment analysis using a Random Forest classifier. In this research, a lot of attention was paid to the optimization of the Random Forest classifier model. This optimization was performed using hyper parameter tuning. This paper described that Random Forest has the following three hyper parameters; number of

features, number of trees and the depth of the tree. This paper also indicated that these hyper parameters must be tuned manually, and described in detail how this process works for each parameter. The following applies to the number of trees, the more trees the better, the stopping point is the number of trees for which the accuracy of the model no longer increases. When selecting the number of features, the selection of random features must be taken into account. The following applies to the depth of trees, the smaller the better, but under fitting must be taken into account. Parmar et al.[29] concluded that Random Forest was an excellent sentiment classification model, the model was found to have a 91 percent accuracy score. The already mentioned sentiment models have yielded good results according to the respective studies. However, one model will have to be chosen for this paper. As a result, these models need to be evaluated. Hossin et al.[15] describes the various evaluation metrics; recall, precision, accuracy and f1 score. These evaluation metrics will be discussed in chapter 5.2.

Chapter 3

Theoretical Background

This chapter describes the different models used for topic modeling and sentiment analysis. The mathematical background behind these models will be presented in this chapter. Topic modeling and sentiment will also be generally introduced.

3.1 Topic modelling

A topic model in machine learning and NLP is a statistical model that is able to distinguish the various topics that appear in the set of documents. This machine learning technique is able to automatically analyze text data and assign this data to a cluster [3]. Topic modeling is therefore an unsupervised machine learning technique; it does not require trained data that has already been classified in advance [3]. Topic modeling is therefore a popular technique that is mainly used in natural language processing, this method can reveal the hidden semantic meanings of the data. As mentioned before, this technique will divide the data into clusters with the same words, these clusters are called the topics [3]. Topic modeling can be performed using various techniques. In this paper, the following three techniques will be applied: Latent Semantic Analysis/Indexing(LSA/LSI), Latent Dirichlet Allocation(LDA) and Hierarchical Dirichlet Process(HDP). These techniques will be discussed below.

3.1.1 Latent Semantic Analysis/Indexing(LSA/LSI)

Latent semantic analysis(LSA) is a statistical model that is able to determine the semantic word similarity between text data. Latent semantic analysis is also called latent semantic indexing, this is due to the purpose of the method, which is indexing text. The aim of this method is to improve the effectiveness of the matching of the semantic value of words [11] . This is achieved through query semantic matching instead of direct word matching. This makes it possible to recognize synonyms and thus assign them to the same semantic meaning. The idea behind this method is that there is a latent structure in the pattern of words usage across documents. This method also assumes that statistical techniques can be used to approximate this latent structure. However, this takes a number of steps.

Before this technique is able to analyze the text, the LSA method creates a matrix with the occurrences of each word in each document. The next step is that LSA uses singular value decomposition (SVD). SVD splits the original matrix into three smaller matrices, which after multiplication are equal to the original matrix. This is called the decomposition step. After this step, the three matrices are again further reduced in size, this is accomplished by choosing a smaller number of dimensions [5]. This process is is shown in figure 3.2.

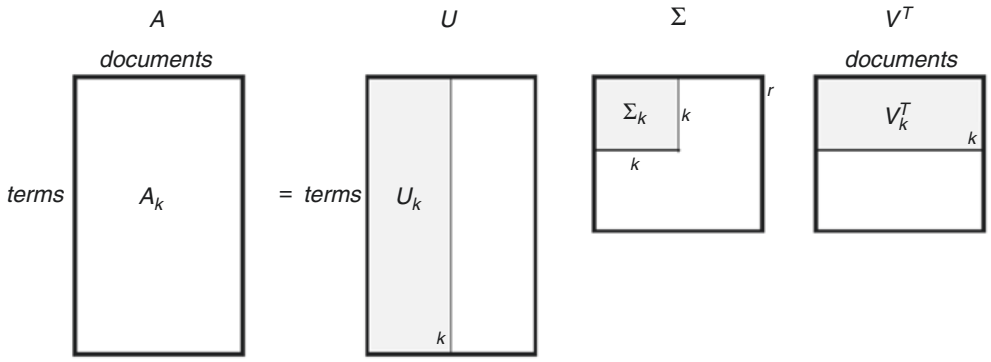


Figure 3.1: Singular value decomposition process in latent semantic analysis Based on [22].

In the figure, A is the original matrix, also called the term-document matrix, U is the left singular vector of words and V is the right singular vector of documents and Σ is a weight matrix. This gives the following formula:

$$A = U \times \Sigma \times V^T. \tag{3.1}$$

The rank of the matrix is r; this method will approximate the original matrix during the dimension reduction (from r to k) [5]. An advantage of this method is that semantic matching is still possible even for documents with no words in common.

3.1.2 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation(LDA) is a probabilistic mixture model, which is used to find latent topics in text data. The word Latent in the name LDA shows that this model tries to find hidden topics from the documents. While the word Dirichlet indicates that the topics in the documents and the words in the topics have a Dirichlet distribution. Allocation refers to the distribution of topics in a document [8]. This model assumes that each document contains a number of different topics and that the words in these documents are generated by these topics. Another assumption of this model is that each document has a different topic distribution [35]. The idea behind this model is that the words in the documents help determine the different topics that appear in the document. Each word in the document is assigned to its topic. This allocation is determined using conditional probability estimates. This process is shown in the figure below.

After the probabilities per word are known, the words must be assigned to the different topics. This can be done in two different ways. The first way is to set a certain

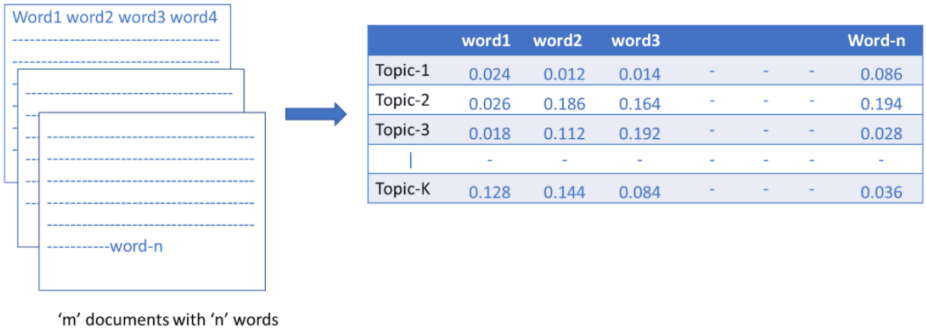


Figure 3.2: Topic probabilities for each word. [35]

probability threshold. As a result, only words with probability equal to or greater than this threshold will be assigned to the corresponding topic. The other way is to say the top x probabilities of words are assigned to the corresponding topic [35]. Determining how likely a document d_i is related to a topic t_k is done using the following formula:

$$p(t_k/d_i) = (n_{ik} + \alpha)/(N_i - 1 + K \times \alpha), \tag{3.2}$$

where n_{ik} is the total number of words in the i-th document that are related to the k-th topic, K is the number of topics, N_i is the number of words in the i-th document, and α is a hyperparameter.

Moreover, determining how likely a word w_j is related to a topic t_k is computed by the following formula:

$$p(w_j/t_k) = (m_{jk} + \beta)/(\sigma m_{jk} + V\beta) \tag{3.3}$$

where m_{jk} is the assignment of the j-th word (W_j) to the k-th topic, V is the Vocabulary of the corpus, and β is a hyperparameter.

3.1.3 Hierarchical Dirichlet Process (HDP)

Hierarchical Dirichlet process(HDP) is the last topic model that has been applied. HDP is a model that clusters grouped data using a Bayesian approach. Like LDA, HDP uses a Dirichlet process for each group of data, meaning that the entire data has the same distribution, which is the Dirichlet distribution. An advantage of this method is the statistical strength that arises, because clusters contain data that belong to several groups of data. HDP shows many similarities with the LDA method, HDP is after all an extension of the LDA method. However, HDP has the great advantage that the number of topics do not have to be clear in advance. The disadvantage of this method is that it is difficult to apply, especially for projects where the number of topics do not necessarily have to be unbounded [38].

3.1.4 Selection Criteria for Topic Models

The above topic models should be compared. However, the question is: how to evaluate topic models? Topic models can be evaluated in three different ways.

Coherence scores	Valuation
0.3	Bad
0.4	Mediocre
0.5	Sufficient
0.65	Great
> 0.65	Unrealistic

Table 3.1: Coherence scores

- The first way is using human judgement. This can be observation based, namely by looking at the top N words in a topic. However, it can also be interpretation based, namely by looking at the words that do not belong to a topic, this is also called topic intrusion [12].
- The second way is using the following quantitative metrics: perplexity or coherence calculations. Perplexity is a measure for comparing probabilistic models. Perplexity is an indication of how well a probabilistic model is able to predict a sample. Usually, the lower this perplexity value is (around 0) the better the topic model functions. Coherence is a measure for the semantic similarity between words in topics. The coherence score varies between 0 and 1, but generally the higher the coherence score, the better the topic model. The coherence scores and the respective valuation are shown below [12].

  These coherence scores are computed for a topic T using the following formula [2]:

  $$Coherencesim(T) = \frac{2\Sigma Sim(W_i, W_j)}{n}, \tag{3.4}$$

  where $Sim(W_i, W_j)$ is the Similarity between the i-th and the j-th words, and n is the number of words in a topic.
- The 3rd way is the combination of human judgment and quantitative metrics.

3.2 Sentiment Analysis

A sentiment analysis is a process in which it is determined for textual data whether the intent of the data is positive, negative or neutral. Sentiment analysis is a natural language processing technique and is one of the most well-known text classification tools [17]. Sentiment analysis is used on a regular basis in business, through this technique companies can gauge customer opinions on various topics such as user satisfaction or the affordability of a product. This is why sentiment analysis is also called opinion mining [19]. This technique enables companies to provide a more appropriate service to their customers. A sentiment analysis can be carried out by means of various classification methods. The classification methods used for this paper are described below, namely Multinomial Logistic Regression, Naive Bayes(TextBlob), Decision Trees and Random Forest.

3.2.1 Multinomial Logistic Regression Classifier

Multinomial Logistic Regression is an extension of the logistic regression method. Logistic regression was only able to classify data into two classes, for example positive and negative or pass and fail. Multinomial logistic regression has the ability to classify the data into multiple classes. This is a major advantage of this method over logistic regression [6]. Multinomial Logistic Regression uses a set of predictor values to classify data into multiple classes. The model determines the probabilities for all possible outcomes for a dependent variable, given that a set of independent variables is used [6]. However, this method has an assumption that must be met before it can be used. This assumption is as follows; the dependent variable must be an ordinal variable or a nominal variable. A nominal variable is a variable that contains several classes and where the order of these variables is not important. Ordinal variables, on the other hand, also have multiple classes, but the ranking of these variables among themselves is in this case very important [34]. Multinomial Logistic Regression is a method with more advantages than just multiclass classification. This method provides good insight into the mutual relationships of variables in a dataset. This method also has a smaller standard error for the parameter estimates than the logistic regression method [34].

3.2.2 Naïve Bayes Classifier

Naive Bayes classification is a supervised learning algorithm; this classification method mainly classifies textual data into various classes. This classification method is known as a fast, simple and accurate technique. The classification is done using the Bayes theory, hence the 2nd term in the name is Bayes. Naive, on the other hand, stands for the fact that this method assumes that the occurrence of a feature is completely independent of another feature. So this is naive [32]. The Naive Bayes classifier is a probabilistic classifier; this means that conditional probabilities determine which object belongs to which class. A conditional probability is a probability determined using prior knowledge. The formula on which this classifier is based is shown below [32].

$$p(A/B) = \frac{p(B/A) * p(A)}{p(B)}, \tag{3.5}$$

where p(A/B) is the posterior probability, $p(B/A)$ is the likelihood probability, $p(A)$ is the prior probability, and $p(B)$ is the marginal probability.

In short, Naive Bayes classification has both advantages and disadvantages. Naive Bayes is a fast and easy classification technique; it is also capable of performing multi class classification in addition to binary classification. However, the major disadvantage is that this method assumes that the features are completely independent of each other, which is of course not the case in the real world.

3.2.3 Decision Trees Classifier

Decision Trees is a supervised machine learning technique, best known for being used for classification. This classification method is an insightful method due to its tree structure. In this tree, the internal nodes represent the features of the dataset while the branches represent the decision rules. In addition, the leaf nodes represent the results. Classifying data using decision trees involves a number of steps. The algorithm

starts at the root node, this root node will be compared by the algorithm with a record attribute, after this the algorithm will determine which node will be the next one. In this new node this process will be repeated, this process will be repeated a number of times until the leaf node of the tree is reached. An advantage of this method is that it is very clear and corresponds to how people actually classify things. As mentioned before, a tree structure makes this classification technique more understandable [4].

3.2.4 Random Forest classifier

Random Forest is a supervised machine learning technique, best known for being used for classification. This classifier is based on a concept called: ensemble learning. Ensemble learning is the process of combining multiple classifiers with the aim of solving a complex problem and increasing the final performance of the model. A Random Forest model combines multiple decision trees where it ultimately takes the average, with the aim of increasing the accuracy of the model. In general, the more decision trees, the better the predictive accuracy of the model. There are not only advantages to this classification method, there are also some assumptions that the Random Forest model must meet. The correlation between the predicted values of each tree must be very low. In addition, this model also assumes that the feature variable in the dataset consists of actual values [4].

Chapter 4

System Architecture

The purpose of this paper is to determine the sentiment per value, but a number of steps are required to achieve this result. This chapter describes the various different research steps that have been carried out. These research steps are schematically represented in an overall workflow. This can be seen in figure 4.1. The remainder of this chapter will focus on the specific techniques per research step.

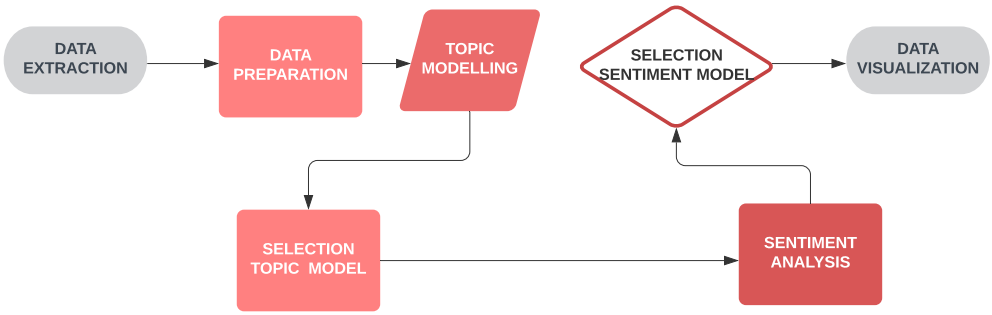


Figure 4.1: Overall workflow

4.1 Data extraction

The first step of the research was to obtain the data. In this study, data was mined for two wireless mobile companies, namely Mint Mobile and Infinity Mobile. This data was mined from Twitter using Twint. Twint is an advanced python tool and scrapes tweets using keywords or hashtags. In this study, the tweets were mined using the keywords: Mint Mobile and Infinity Mobile. Twint is a method that scrapes tweets without an API. The advantage of this is that there is no longer a limit on the maximum number of tweets that can be mined. This research step has thus resulted in two different data sets, which contained 67k tweets in total. These tweets have been extracted for the period 1 January 2019 to 1 June 2021, a period that coincides with the pandemic but also includes a period before COVID-19. Due to the size of the data, the data was mined in steps of three months, and the data was merged afterwards. The data

contains the following relevant variables; the userid (the unique code of the user), the tweetid (the unique code of the tweet), the date (date of the tweet), the tweet, the retweet (whether it is a retweet or not), the language (the language of the tweet), nlikes (the number of likes of the tweet) and the hashtags (of the tweet). The tweets have an average character length of 80. The distribution of the character lengths and the word clouds of the data can be seen in Appendix A.

4.2 Data preparation

The first step of the data preparation is the data cleaning; the dataset must be cleaned before a data analysis can be performed. The cleaning has been done in several steps. For the research it was relevant to only use English tweets. The first data cleaning step was therefore to remove all non-English tweets from the dataset. The 2nd step was deleting the retweets. In the dataset, some tweets turned out to be retweeted very often, which could lead to distorted results in the sentiment analysis. An added benefit of deleting retweets is that the chance of bots also decreases, bots retweet more often than real accounts. The 3rd step was a checkup step. In this step, the tweets were checked for uniqueness based on the tweetid. This step shouldn't shrink the dataset, otherwise something would have gone wrong when deleting the retweets.

The data is now cleaned but the data must be fully prepared before it can be used for topic modelling. After the cleaning there will be several steps to complete the data preparation phase. The first step is to adjust the format of the tweets, so that the result of the topic model will be better. The tweets will therefore be converted into a BOW (bag of words) corpus. In this corpus, the tweets will be split into a list of words containing each tweet. The order of the words in such a list is not important for a bag of words corpus. The next step is essential for the format change described above to be successful. In this step, hyperlinks, punctuations and numbers that appear in the tweets will be removed. This results in a clean list of words per tweet. The next step is to remove the stop words from this list. This step is very important as stop words like "the", "is" or "on" have no semantic meaning and are therefore not relevant for both the topic model and the sentiment analysis. As a result, the quality of the topics will be higher and we will not have topics with the same top words. Finally, the upper case words must also be converted into lower case words.

The last step is to remove outliers, for example words that barely occur in the corpus or always occur (stop words). Words that barely occur are, for example, names of twitter users or specific language that is only known in the informal circle. Reducing this noise is done with the help of two hyper parameters namely MINDF and MAXDF. MINDF is a threshold value; the minimum number of times a word must appear in the corpus in order to not be deleted. While MAXDF stands for the maximum number of times a word may appear before it is removed from the corpus. Because the stop words have already been removed from the corpus, the MAXDF is set to 1, which means that a word is only removed if it occurs in all documents. The MINDF has been set to 20 after much experimentation. Experimentation has led to an optimal threshold. The entire data preparation process is summarized in figure 4.2.

Figure 4.2: Data preparation process [1].

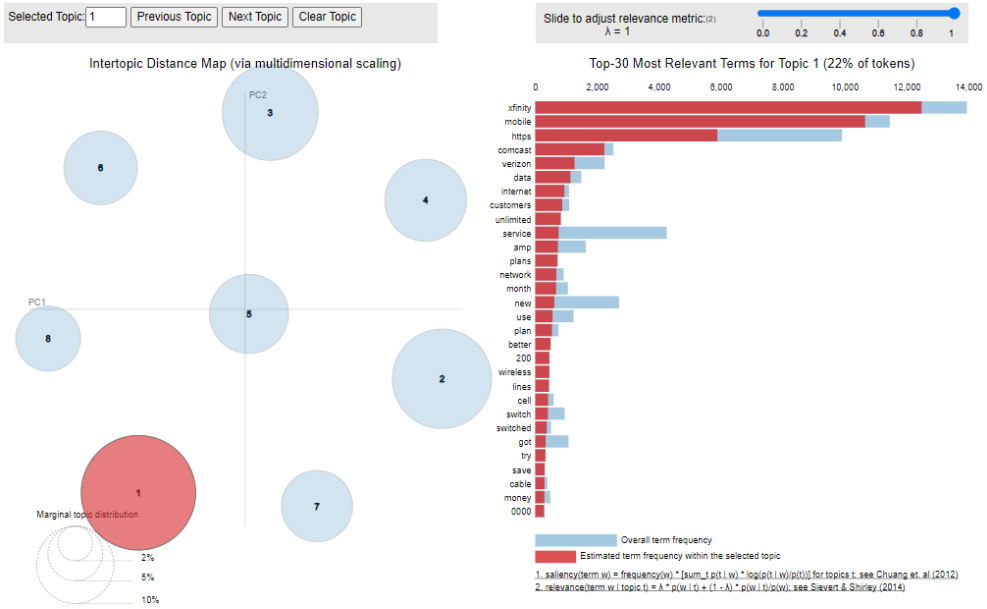
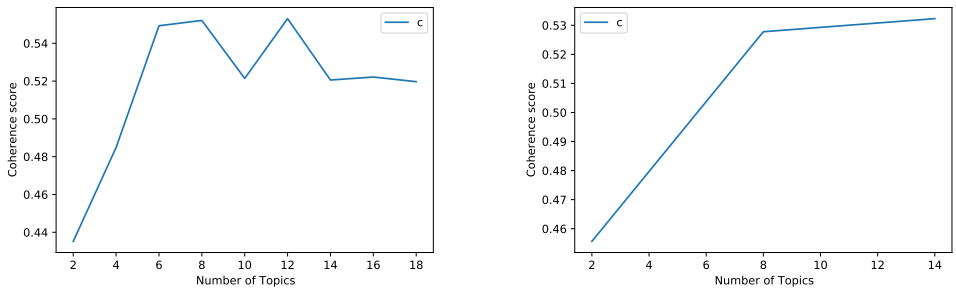


Figure 4.3: Topic modelling visualization user satisfaction for Mint Mobile

4.3 Topic Modelling

After the data extraction and preparation, the data is ready for the topic modeling step. In this step three different topic models were built namely Latent Semantic Analysis(LSA), Latent Dirichlet Allocation(LDA) and Hierarchical Dirichlet Process(HDP). The operationalization of these models were already discussed in the subchapters 2.1.1, 2.1.2 and 2.1.3, however the topic modeling process will be described in this step. This process is the same for all topic models and is realized using the python library pyLDAvis. Using this library is the first step in the topic modeling process. This library visualizes the different topics in a figure, which consists of several circles. Where each circle must represent a topic, the further the circles are from each other, the more different these topics are. The size of these circles gives an indication of the amount of data that belongs to this topic. PyLDAvis also visualizes the 30 most relevant terms for each topic, whereby for each term the estimated term frequency within the topic is displayed in relation to the overall term frequency. PyLDAvis visualizes this by means of a histogram on the right side of the figure. This makes it possible to read per topic which words specifically belong to the relevant topic. This makes it possible to see per topic what is being discussed and to assign this topic to a category. The next step is to perform this process for the values user satisfaction, affordability and willingness

for each company separately. Figure 4.3 shows the visualization for the Mint Mobile dataset for the value user satisfaction. The visualizations for the values affordability and willingness are shown in Appendix B. The last step in the topic modeling process is to assign the corresponding topic to each tweet in the dataset. This gives you 3 different datasets for each company dataset. Examples of tweets associated with user satisfaction, affordability or willingness are shown in Appendix C.



(a) (b)

Figure 4.4: Optimal number of topics estimated using (a) LDA (b) LSI. In the figure legend, c refers to the cohernece score.

4.3.1 Selection topic models

In this step, the three topic models will be evaluated. This step will consist of two sub steps. The first sub step will be the optimization of each topic model using hyper parameter tuning. There are two hyper parameters that directly influence a topic model, namely the number of topics and the number of iterations. However, the HDP model automatically extracts the number of topics from the data, so the number of topics is not tuned for this model. The number of topics are very essential, as too few topics will result in existing topics being ignored/hidden. Too many topics can lead to wasted topics, topics that belong together will be separated. In this research two ways were used to analyze the number of topics; manually and quantitatively by means of coherence scores. In the first method, the results of the topic models for the range of 5 to 15 topics were manually examined. The result was assessed on the meaning of the words in each topic and the weight of the words. This analysis had shown that 8 to 9 topics gave the best results. The next step was the quantitative analysis in which the coherence score was determined for each number of topics. The cut-off point (elbow method) for the coherence score turned out to be at 8 topics. The corresponding graphs can be seen in the figures below. This study therefore showed that the best results were obtained for 8 topics.

The number of iterations was determined by means of manual experiments in which a number of iterations were chosen within the range of 50 to 1000. This experiment showed that there was hardly any difference in the result if the number of iterations were between 100 and 200. As a result, 150 iterations were chosen for this study. The last sub step was comparing the models with each other, based on the coherence scores. The model with the highest coherence score has the highest semantic similarity within the topics. This model will be selected as the final topic model. The evaluation showed

that the LDA model with a coherence score of 0.55 gave the best result, followed by the LSA model with a score of 0.52 and finally the HDP model with a score of 0.19. As a result, the LDA model was chosen as the topic model in this study. The coherence scores for each model are summarized in a plot in Appendix A.

4.4 Sentiment Analysis

In this step the different sentiment models will be trained. In this research, the following sentiment models will be discussed: Multinomial Logistic Regression, Naive Bayes, Decision Trees and Random Forest. Multinomial Logistic Regression, Decision Trees and Random Forest are models built in the same way. For these models, a dataset was used with tweets of which it is already known in advance whether these tweets were positive, neutral or negative. This dataset was used to train the models. This dataset was split into a training dataset and a test dataset, with a ratio of 4:1. The test dataset was used to map the performance. The next step was vectorising the dataset with a BOW (bag of words) model. The first step to achieve this was to create a dictionary, which is a list of unique words. After that, it was possible to describe each document as a vector with words in the dictionary. The next step was to determine the frequency of each word in a document, which was accomplished using the TF-IDF step (term frequency, inverse document frequency). This also determines how distinct a word is compared to other documents. After this, the appropriate classifier classified the test data tweets. Now that the relevant classifier was trained, it was possible to start with the last step. The last step was to classify the tweets of the research dataset. The Naive Bayes model deviates from this step-by-step plan because Text blob is an unsupervised learning method. Text blob is able to calculate a polarity score for each tweet in the dataset. The polarity scores range from -1 to 1, with anything below zero being classified as negative and anything above 0 as positive. If the polarity is exactly 0 then the tweet is neutral. An advantage of this classifier is that it works with polarities, which gives some insight into the extent a tweet positive or negative is.

4.4.1 Selection sentiment model

Prior to this step, the different sentiment models were built. In this step, the sentiment models will be evaluated on the basis of a number of parameters. These parameters are: accuracy, recall, precision and f1 score. However, a sub step is required to achieve this result. In this sub step, the various sentiment models will first have to be optimized, using hyper parameter tuning. For the Multinomial Logistic Regression model the solver hyper parameter is set to 'sag' while for logistic regression the default solver is 'lbfgs'. The reasoning behind this is that with multi classification, 'lbfgs' leads to multinomial losses and therefore a less accurate model. The penalty parameter for a Multinomial Logistic Regression model is set to 'L2', this parameter prevents under fitting. For a Random Forest and Decision Tree model, the complexity of the model plays a very important role. The hyper parameters that needed to be tuned directly depended on it. Hyper parameters such as the depth of the trees, the maximum amount of trees and the number of features. The maximum number of trees has been determined manually by testing a range from 50 to 400 trees and looking at the accuracy of the model. This experiment showed that after 100 trees the accuracy no longer changes. This is also known as the stop criterion. The maximum depth and

number of features were tested using exactly the same method, but the default values showed the best results. For the Naive Bayes classifier (Text blob) no hyper parameter tuning was applied as this model was totally unsupervised. In the last sub step, a classification report was drawn up for each model, in which the parameter values were shown for each sentiment. The closer the parameter values are to 1, the better the score. In addition, a confusion matrix was also drawn up for each model. This matrix showed how many tweets from the test data set have been classified correctly and how many tweets have been classified incorrectly. The confusion matrices and the original classification reports are shown in the code. Finally, on the basis of the average accuracy, the final value judgment was given. The results of the evaluation and the evaluation steps in detail will be discussed in the next chapter.

4.5 Data Visualization

The previous steps made it clear which sentiment model and topic model perform best. The next step in the system is therefore the visualization of the results. However, this again takes place in several steps. The first step is to visualize the sentiments per company, so that these companies can be compared with each other. The second step is to visualize the sentiments for each of the values. The entire dataset will be used for this step. This step will provide insight into the extent to which there are differences in sentiment between the values; user satisfaction, affordability and willingness. The last step is to provide insight into the sentiments over time using a time analysis. This step will consist of two sub steps. The first sub step is the weekly mean sentiment values over time, this analysis aims to determine the trend. It can also be concluded from this analysis whether COVID-19 played a role. The 2nd sub step is the monthly mean sentiment values over time, this analysis aims to get a better view of the months that deviate from the average sentiment values. This analysis will also map the transition between pre-COVID and COVID better.

Chapter 5

Evaluation

This chapter will describe the two data sets used to evaluate the system, the evaluation metrics used to gauge the performance of the system will also be discussed. The applied baseline methods will also be explained in this chapter. In addition, the models will be compared and a comparison will be made with the literature. Finally, the results of the analysis will be presented.

5.1 Datasets

To determine and evaluate the prediction accuracy of the different models, two different datasets were used. These are datasets that are regularly used for research papers on sentiment analysis. The two datasets are; the Twitter US Airline dataset and the tweet sentiment extraction dataset. The choice for these datasets is based on both the domains and the size of the datasets. The Twitter US Airline is a dataset ideal for gauging sentiment around a service or product, while the tweet sentiment extraction dataset is a COVID-19 related dataset. These domains both fall within my research; despite they differ from each other. There is also a size difference between the datasets, the US Airline dataset is the smaller dataset while the COVID-19 dataset is the larger dataset. The contrasts between this dataset will deepen the evaluation of the different models.

The Twitter US Airline dataset is a dataset released by CrowdFlower. It is a dataset containing 14640 tweets for 5 major US airlines; Delta, Virgin America, Southwest, United and US airways. These tweets are pre-labeled negative, neutral and positive. The dataset was scraped in February 2015 from Twitter, since then, this dataset has become one of the most well-known sentiment datasets. The tweet sentiment extraction dataset is a dataset released by Kaggle for a Kaggle competition. It is a dataset containing 27481 COVID- related tweets. These tweets are pre-labeled negative, neutral and positive. The dataset was scraped in 2020 from Twitter, since then, this dataset has been used frequently by competitors aiming to win the 15,000 dollar prize money.

Before the datasets were ready for use, there were a number of data preprocessing steps. The first step was to remove the columns without data for the purpose of

Datasets	Twitter US airline sentiment	Tweet sentiment extraction
Domain	US airlines	COVID-19
# of attributes	8	5
# of instances	14640	27481
# of sensitive attributes	1	1
Name of the sensitive attribute	tweet	tweet
Variables	Tweetid, sentiment, label, negativereason, negativereason_c, airline, name, retweet_count	Tweetid, tweet, selected text, sentiment, label
Decision labels	Negative (=0), Neutral(=2) and Positive(=4)	Negative (<0), Neutral(=0) and Positive(>0)
Binary labels	No	No

Table 5.1: Dataset details

dimensionality reduction. The next step was to delete the rows with missing data. The last step was to homogenize the decision labels. The Twitter US Airline dataset linked different values to the different sentiments than the tweet sentiment extraction dataset. Ultimately, the values of the tweet sentiment extraction dataset were taken as the yardstick, i.e. negative (between -2 and 0), negative (=0) and positive (between 0 and 2). Lastly, both datasets have one sensitive attribute and have multiclass decision labels. The details for each dataset are described in Table 5.1. The variables mentioned in the table are the variables that contain data. The completely empty columns are therefore not listed in this table. The values listed in the rest of the table are the values associated with the data before preprocessing started.

5.2 Evaluation metrics

Several classifiers have been applied on the datasets, these classifiers should be assessed on their functioning. How these models are evaluated will be discussed in this subchapter. There are several metrics that determine how good the predictions are of your sentiment classifier model. These metrics are the accuracy, precision, recall and F1 score. However, before these metrics can be determined, the terms True Positive(TP), True Negative(TN), False Positive (FP) and False Negative (FN) must be defined. True Positive(TP) stands for the positive values that have been correctly predicted. This means that a tweet that is positive has actually been recognized as positive by the model. True Negative(TN) stands for the negative values that have been correctly predicted. This means that a tweet that is negative has actually been recognized as negative by the model. False Positive(FP), on the other hand, represents a positive prediction while the actual class is negative. This means that a tweet that is negative has been recognized as positive by the model. Hence the term false positive. False Negatives (FN) is the exact opposite, so a tweet that is positive is recognized as negative by the model [28]. The confusion matrix is a matrix that can map the TP, TN, FP and FN well. This matrix will give you a better understanding of what these terms mean. Table 5.2 shows this confusion matrix.

The accuracy is one of the most well-known evaluation metrics. This metric measures

	Actually Positive(1)	Actually Negative(0)
Predicted Positive(1)	True Positives(TPs)	False Positives(FPs)
Predicted Negative(0)	False Negatives(FNs)	True Negatives(TNs)

Table 5.2: This table shows the confusion matrix. This matrix shows the performance of a classifier on test data for which the actual values are already known. [15]

the ratio between the correctly classified observations, so the True Positives and True Negatives and all classified observations. Generally, the higher this accuracy, the better the model. However, there is a caveat to this metric, this metric works best with symmetrically distributed datasets. This means that the ratio of False Positives and False Negatives must be around 1. This means that other metrics should also be considered in combination with the accuracy metric for a better insight into how a model works [15]. The formula for computing the accuracy is:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}. \tag{5.1}$$

The precision, on the other hand, is a measure of the amount of correctly classified positive observations relative to all positive predictions. This metric says something about how well positive tweets are recognized by the classification model [15]. The higher the value of this metric, the better the model is. The formula for precision is shown below.

$$Precision(P) = \frac{TP}{TP + FP}. \tag{5.2}$$

The recall or sensitivity is yet another evaluation metric. This metric is a measure of the amount of correctly classified positive observations relative to all positive observations. This metric therefore determines the ratio between the positively labeled observations and the number of actual positive observations. If this metric value is higher than 0.5, it is a good score [15]. The formula for the recall is shown below.

$$Recall(R) = \frac{TP}{TP + FN}. \tag{5.3}$$

The F1-score is an evaluation metric that takes the average of the recall and the precision. As a result, this metric will include the False Negatives and False Positives in the evaluation. It also has the great advantage that the dataset does not have to have an even class distribution as with the accuracy. This makes the F1-score one of the most important evaluation metrics [15]. The formula for the F1 score is shown below.

$$\text{F1-score} = \frac{2 * R * P}{R + P}. \tag{5.4}$$

5.3 Baseline methods

Before the different models can be evaluated, a baseline model must be built. This is a simple model that aims to make the classification process more transparent. For example, a baseline provides insight into the classes that are more difficult to distinguish or certain aspects of the data that are missing from the model[18]. In this study,

a baseline model was built for each dataset using the dummy classifier. The dummy classifier classifies data using a few simple rules, resulting in a baseline performance. The baseline performance is the prediction accuracy that is achieved by betting the correct classes. The intention of a baseline performance is that the other classifiers should deliver a better performance on the same dataset. So it serves as a lower bound. These baseline methods have a very good effect on unbalanced datasets, which shows the benefit of using a baseline. As indicated earlier, the baseline depends on simple rules that are also called strategies. The operation of the classifier is therefore independent of the training data[18]. The dummy classifier has a number of strategies, in this research the following two will be explored; stratified and uniform.

- Stratified is a method where the predictions follow the classes distribution in the training data.
- Uniform is a method in which the predictions of the classes are totally uniformly random.

These baseline performances will be compared in chapter 5.5.

5.4 Experimental setup

The experimental setup of the model evaluation will consist of several steps. These steps are listed below.

- The first step is to compare the baseline results. This comparison will be between the two different strategies but also between the different datasets. The highest performance of the baseline model will be the lower limit when comparing the more complex models.
- In this second step, the performances of the classifiers; Multinomial Logistic Regression, Naive Bayes, Decision Trees and Random Forest; are compared using the previously mentioned evaluation metrics.
- In the last step the performances of the classifiers; Multinomial Logistic Regression, Naive Bayes, Decision Trees and Random Forest; are compared with the performances of the same classifiers in an already existing research paper.A research paper that has built a similar system.

These steps will result in a final model that will be used for the analysis.

5.5 Comparison of the models

Table 5.3 shows the baseline results for each strategy and for each dataset. The values in this table are the weighted averages for the precision, recall, F1-score and accuracy. This table shows that the stratified strategy yields a higher accuracy for both datasets than the uniform strategy. This is in line with the theory since the uniform strategy assumes that the predictions of the classes are totally uniformly random, while with the stratified strategy the predictions follow the classes distribution in the training data. However, the difference in the accuracy between the strategies for the tweet sentiment extraction dataset is small. The maximum accuracy for the baseline models are 0.47 and 0.35, these values are an evaluation benchmark for the more complex

Dataset	Splitting	Precision	Recall	F1-score	Accuracy
US airline dataset	Stratified	0.47	0.47	0.47	0.47
	Uniform	0.47	0.32	0.36	0.32
Tweet sentiment extraction dataset	Stratified	0.35	0.35	0.35	0.35
	Uniform	0.34	0.33	0.34	0.33

Table 5.3: Classification performance for the dummy classifier.

Classifier	Sentiment	Precision	Recall	F1-score
Multinomial Logistic Regression	Negative	0.86	0.90	0.88
	Neutral	0.64	0.61	0.62
	Positive	0.76	0.69	0.73
Naive Bayes	Negative	0.88	0.35	0.50
	Neutral	0.32	0.57	0.41
	Positive	0.33	0.76	0.46
Decision Trees	Negative	0.78	0.81	0.79
	Neutral	0.45	0.45	0.45
	Positive	0.60	0.53	0.56
Random Forest	Negative	0.75	0.96	0.84
	Neutral	0.67	0.37	0.47
	Positive	0.82	0.46	0.59

Table 5.4: Classification performance of the framework with different classifiers on US Airline dataset.

models. The more complex ones should have a higher accuracy than 0.47/0.35. It can also be observed that the values for the different evaluation metrics are close to each other. The variance in the baseline values is therefore small.

Table 5.4 shows precision, recall and F1-score for each model. These evaluation values are the result of the classification on the US Airline dataset. From the table, it can be seen that all models have a lower evaluation metric score for a neutral sentiment. Correctly recognizing neutral tweets therefore seems to be a weakness of all models. However, the Random Forest model with a score of 0.37 is the model that has the most difficulty in correctly classifying neutral tweets. In contrast, the evaluation metric scores for a negative sentiment are excellent. In addition, it is striking that the baseline accuracy benchmark is significantly improved by Multinomial Logistic Regression, Decision Trees and Random Forest models, while the Naive Bayes classifier is even slightly below this benchmark. The Naive Bayes classifier is the only unsupervised classifier and shows a worse performance than the supervised classifiers.

The values in the table can be summarized by the model accuracy.The Multinomial Logistic Regression model gives the best result with an accuracy of 0.80, followed by the Random Forest with an accuracy of 0.75 and the Decision Tree model with an accuracy of 0.69. Finally, Naïve Bayes is the worst model with an accuracy of 0.46 (below the baseline benchmark). For this dataset, the Multinomial Logistic Regression model performed the best, for the tweet sentiment extraction dataset, the same models will again be evaluated.

Classifier	Sentiment	Precision	Recall	F1-score
Multinomial Logistic Regression	Negative	0.69	0.62	0.66
	Neutral	0.65	0.73	0.69
	Positive	0.78	0.72	0.75
Naive Bayes	Negative	0.66	0.46	0.54
	Neutral	0.60	0.51	0.55
	Positive	0.54	0.80	0.65
Decision Trees	Negative	0.58	0.58	0.58
	Neutral	0.61	0.63	0.62
	Positive	0.70	0.67	0.69
Random Forest	Negative	0.73	0.54	0.62
	Neutral	0.62	0.79	0.69
	Positive	0.79	0.68	0.73

Table 5.5: Classification performance of the framework with different classifiers on Tweet sentiment extraction dataset.

Table 5.5 shows precision, recall and F1-score for each model. These evaluation values are the result of the classification on the tweet sentiment extraction dataset. From this table it can be seen that the differences in the evaluation metric scores between the models are much smaller than with the US Airline dataset. It can also be seen that the models don't have a very bad evaluation metric score for a specific sentiment. For the US airline dataset, this was the case for the neutral sentiment. In this case, the evaluation metric scores of the various sentiments are closer to each other. The variance in the evaluation metric scores is therefore lower, but it is also striking that the average evaluation scores are lower than with the US airline dataset. In addition, it is also striking that for this dataset the positive sentiment has a higher evaluation metric score than the negative sentiment. In the US Airline dataset, this was the other way around.

The values in the table can be summarized by the model accuracy. The Multinomial Logistic Regression model gives the best result with an accuracy of 0.70, followed by the Random Forest with an accuracy of 0.69 ,the Decision Tree model with an accuracy of 0.63 and then the Naïve Bayes model with an accuracy of 0.59.Another important point is that the baseline accuracy benchmark is significantly improved by all models. So the Naïve Bayes classifier has a much better performance for this dataset than for the US Airline dataset. However, it still remains the worst-functioning sentiment model. Lastly, the difference in accuracy between Random Forest and Multinomial Logistic Regression has become smaller for this dataset, but the Multinomial Logistic Regression model again has the best performance.

The final step of the evaluation is to compare the results of the system for this study with similar systems mentioned in research papers. Table 5.6 contains the evaluation metric scores for each dataset. Rane and Kumar[31] investigated the performance of multiple multiclass classifiers using the US Airline dataset. Liu et al. [21] investigated the performance of multiple multiclass classifiers, but then using the tweet sentiment extraction dataset. Table 5.6 shows that for the US Airline dataset the Random Forest model gives the best result with an accuracy of 0.84, followed by the Multinomial Logistic Regression with an accuracy of 0.81 and the Decision Tree and Naïve Bayes

Dataset	Classifier	Precision	Recall	F1 score	Accuracy
US Airline dataset	Multinomial Logistic Regression	0.81	0.82	0.82	0.81
	Naïve Bayes	0.64	0.65	0.64	0.64
	Decision Trees	0.63	0.64	0.65	0.64
	Random Forest	0.85	0.86	0.86	0.84
Tweet sentiment extraction dataset	Multinomial Logistic Regression	-	-	-	0.68
	Naïve Bayes	-	-	-	0.55
	Decision Trees	-	-	-	0.60
	Random Forest	-	-	-	0.63

Table 5.6: Classification performance from [31] for US Airline dataset and from [21] for Tweet sentiment extraction dataset.

models with both an accuracy of 0.64. The Random Forest model has the highest accuracy while for our system Multinomial Logistic Regression had the highest accuracy. The reason for this is that the system described in the paper by Rane[31] does not use topic modeling before the sentiment classification step. This results in the fact that the complete dataset is used for the sentiment classification, and Random Forest models are known for their good performance with larger datasets. It can also be seen that for the tweet sentiment extraction dataset, the Multinomial Logistic Regression model has the highest accuracy, which corresponds to the result obtained for our system. From this three-step evaluation it can be concluded that the Multinomial Logistic Regression model is the best working model for the system. As a result, this model was chosen as the final model with which the analysis was performed.

5.6 Results

In the previous subchapters, the various models were evaluated and the model with the best performance was chosen. This subchapter will display the analysis results achieved using the chosen model. Chapter 5.6.1 will present the average sentiment analysis results for the companies Infinity Mobile and Mint Mobile. Chapter 5.6.2 will display the complete sentiment analysis results for each value. Chapter 5.6.3 contains the time analysis.

5.6.1 Sentiment companies

Figure 5.1 shows the average sentiment scores per company for each value. The sentiment scores range from -2 to 2 with between -2 and 0 being negative, between 0 and 2 being positive and 0 being neutral. Figure 5.1 shows that for user satisfaction, Xfinity mobile has a sentiment score of -1.3 which is a very negative sentiment while Mint Mobile has a sentiment score of -0.6 which is an average negative sentiment. From this plot it can be seen that for affordability, Xfinity mobile has a sentiment score of -1.0, which is a negative sentiment while Mint Mobile has a sentiment score of -0.3

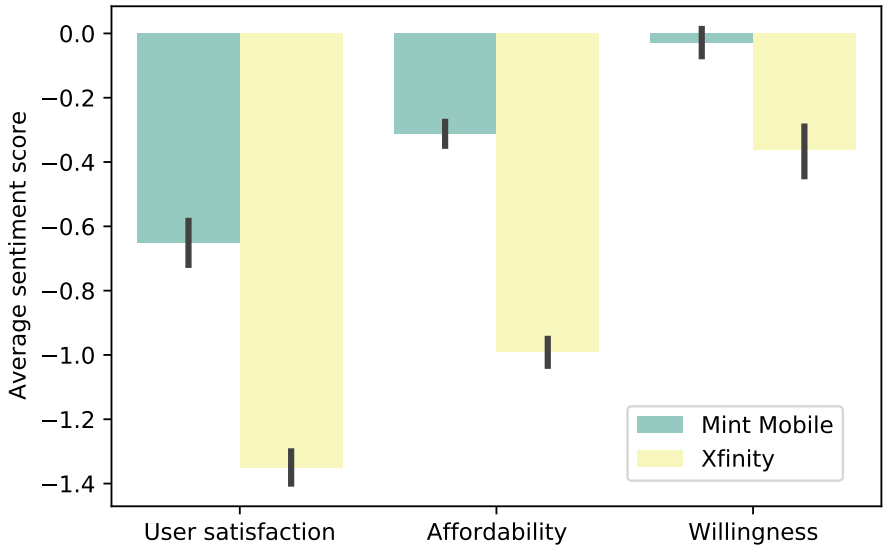


Figure 5.1: Sentiment bar plot companies.

which is a slightly negative sentiment. For the value willingness, Xfinity mobile has a sentiment score of -0.4, which is a slightly negative sentiment while Mint Mobile has a sentiment score of -0.05, which is very close to 0 and can therefore be classified as neutral sentiment. These sentiment plots show that Xfinity mobile has lower sentiment scores than Mint Mobile and therefore a more negative sentiment. This applies to all the different values.

5.6.2 Overall Sentiment Distribution

Figures 5.2 and 5.3 show the overall sentiment distribution for each value. In this analysis, no distinction has been made between the two different companies. The figure shows for each value, how many tweets are classified as negative, neutral or positive. For user satisfaction, by far the most tweets were classified as negative, about 2100 tweets, neutral about 1100 tweets and positive about 250. The negative sentiment is twice as much as the neutral sentiment, while the positive sentiment is negligibly small.

For affordability, the differences between the negative and neutral sentiment are closer. Most tweets for this value are classified as neutral with about 3750 tweets, the negative sentiment has about 3200 tweets and positive 750 tweets. Positive tweets again occur considerably less often than the neutral and negative tweets.

For the willingness, the positive sentiment is quite significant which makes the differences between the sentiments narrower. Most tweets for this value are classified as neutral with about 2500 tweets, the negative sentiment has about 1900 tweets and

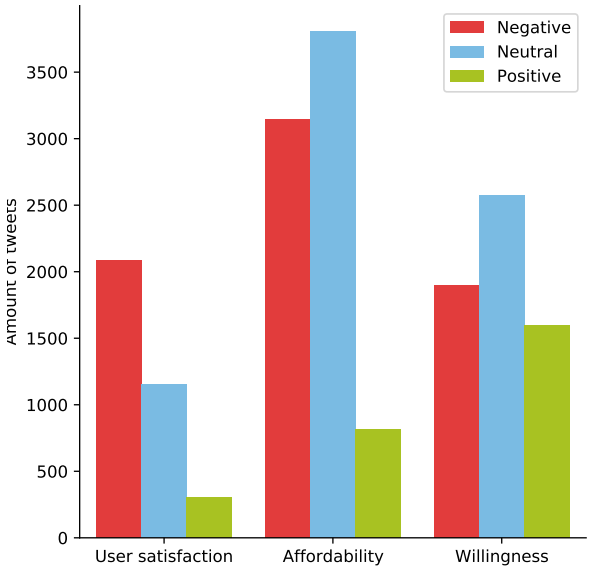


Figure 5.2: Overall sentiment distribution (count).

positive has 1600 tweets. This resulted in the sentiment ratios that can be seen in Figure 5.3.

5.6.3 Time analysis

In the time analysis, the sentiment changes over time were examined separately for each value. The time period for which this was studied was from 01-01-2019 to 01-06-2021, with the aim of determining sentiment and mapping out the possible effects of the pandemic. The analysis was performed by creating a scatter plot with the weekly average values. In this analysis, a second degree polynomial has also been fitted through the data points, making the trend more visible. The starting point of the COVID pandemic has been set at 01-03-2020, this represents the vertical line in the plots. The horizontal line represents the transition line between positive and negative. As mentioned before, the sentiment scores range from -2 to 2 with between -2 and 0 being negative, between 0 and 2 being positive and 0 being neutral.

Figure 5.4 shows the time analysis plot for the value: user satisfaction. The figure shows that the sentiment for user satisfaction is around -1, which is clearly a negative sentiment. It can also be seen that the variance in sentiment over time is very small. The data points are all in the negative sentiment range. This results in an almost horizontal trend line.

Figure 5.5 shows the time analysis plot for the value: affordability. The figure shows that the sentiment for affordability is slightly negative. However, it is clear that sentiment before the pandemic has been trending upwards towards neutral sentiment. Around the start of the COVID pandemic, a clear trend break can be observed. During the pandemic, we see a downward trend towards a sentiment value of -1. The

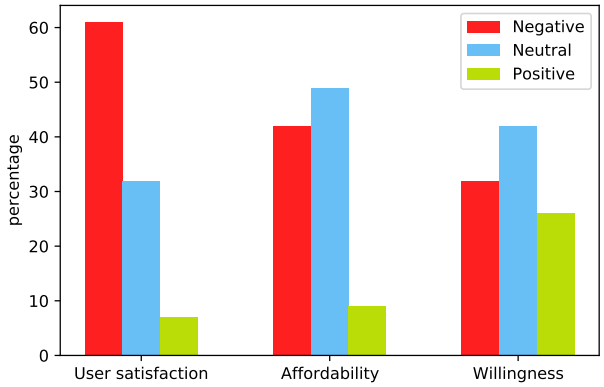


Figure 5.3: Overall sentiment distribution (ratio).

variance of the data points over time for affordability is therefore greater than for user satisfaction.

Figure 5.6 shows the time analysis plot for the value: willingness. The figure shows that the sentiment for willingness was positive before the pandemic and negative after the start of the pandemic. However, it can be clearly seen that pre-pandemic sentiment was already trending downward towards neutral sentiment. Around the start of the COVID pandemic, a clear trend break can be observed. The downward trend just before the pandemic initially appeared to be decreasing, but at the start of the pandemic, a downward trend is set again towards a sentiment value of -1. The variance of the data points over time for willingness is therefore very large, the sentiment ranges from 1.25 (very positive) to -1.25 (very negative).

In Appendix C, the corresponding monthly time analysis figure is shown for each value. The monthly time analysis aims to map out the outliers. For user satisfaction, as with the weekly figure, it can be seen that there is a negative sentiment with very small variance. However, there is one month that can be seen as an outlier, namely March 2019. This month has a sentiment score of -1.6 while the average sentiment score is -1. In addition, it can also be seen that for user satisfaction, the transition from pre-COVID to COVID has no influence on the sentiment score.

As with the weekly figure, the value affordability shows a negative sentiment. A few months stand out from this figure, these months can be labeled as an outlier. March 2019 and January 2021 stand out because they have a very negative sentiment score. These months have a sentiment score of -1.0 while the average sentiment score is -0.6. Another month that stands out is December 2019, this month stands out because of its less negative sentiment. This month has a sentiment of -0.3 while the average is -0.6. In addition, it can also be seen that for affordability, the transition from pre-COVID to COVID does indeed influence the sentiment score. The months surrounding the transition have a less negative sentiment score as both the months before and after the transition months.

As with the weekly figure, the value willingness also shows both negative and positive sentiment. A number of months stand out from this figure, which can be labeled as an outlier. April 2021 is notable for a very negative sentiment. This month has a sentiment score of around –0.8 while the average is 0.0. Another month that stands out is January 2019, this month stands out because of its positive sentiment. This month has a sentiment of 0.8 while the average is 0.0. In addition, it can also be seen that for willingness, the transition from pre-COVID to COVID does indeed influence the sentiment score. The months surrounding the transition have exactly a sentiment of 0.0 and are exactly the months in which the sentiment changes from positive to negative.

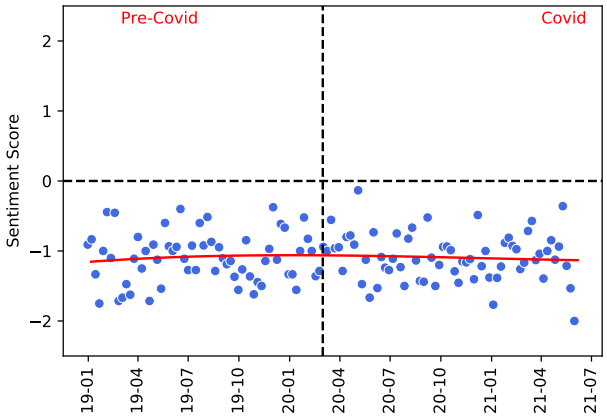


Figure 5.4: Time analysis user satisfaction

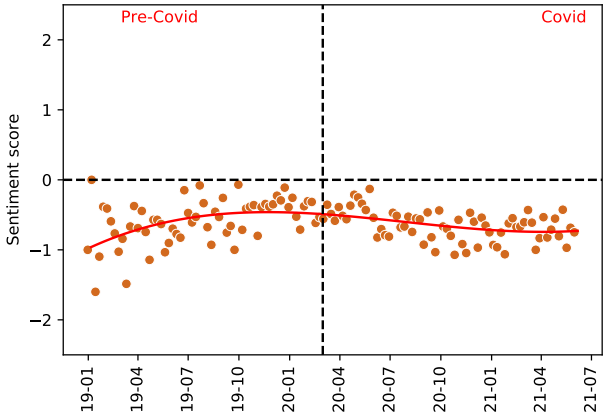


Figure 5.5: Time analysis affordability

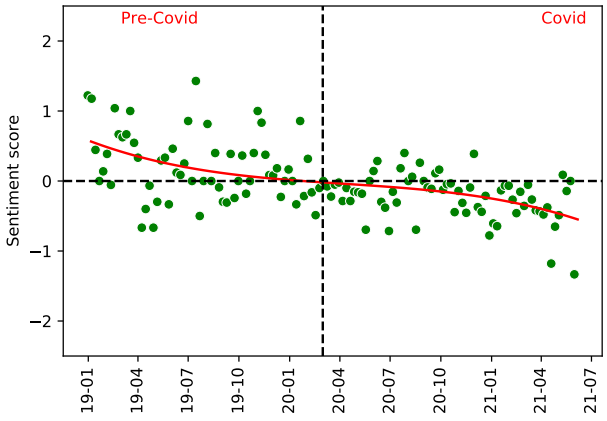


Figure 5.6: Time analysis willingness

Chapter 6

Conclusion

In this chapter, the summary, answers to the research questions, the limitations of the research, ethical considerations and further work will be discussed.

6.1 Summary

Wireless mobile services have become the fastest growing part of the telecommunications sector. The use of a mobile phone has become an essential part of today's society. Today, 90 percent of the world's population over the age of six has access to a mobile phone, i.e. billions of people. This makes wireless mobile communication a worldwide phenomenon, for developed countries as undeveloped countries. The rise of mobile technology therefore has a direct social-economic impact. Therefore, this paper examines the social-economic impact of wireless mobile services for the companies Infinity Mobile and Mint Mobile. These two companies contradict each other in their way of delivering services. This makes it interesting to investigate whether a company's policy influences the sentiment of core values such as user satisfaction (social effect), affordability (economic effect) and willingness (social effect). The research question for this paper is therefore: "How does wireless mobile services impact values as user satisfaction, affordability and willingness?"

The analysis for this research is performed by a created system that consists of several components. Components such as data extraction, data cleaning, data preparation, topic modeling and sentiment analysis. In the data preparation phase, the data is converted into an appropriate format called the BOW (bag of words) corpus. In addition, the hyperlinks, punctuations and numbers that appear in the tweets are removed. Also, the upper case words are converted into lower case words. Outliers are also removed; these outliers are words that barely occur in the corpus or always occur (stop words). Words that barely occur are, for example, names of twitter users or specific language that is only known in the informal circle. Reducing this noise is done with the help of two hyper parameters namely MINDF and MAXDF. MINDF is a threshold value; the minimum number of times a word must appear in the corpus in order to not be deleted. While MAXDF stands for the maximum number of times a word may appear before it is removed from the corpus.

In the topic modeling phase, three different topic models are built namely Latent Semantic Analysis(LSA), Latent Dirichlet Allocation(LDA) and Hierarchical Dirichlet Process(HDP). These models are optimized using hyper parameter tuning. The two hyper parameters that directly influence a topic model are the number of topics and the number of iterations. Finally, the best working topic model is selected on the basis of the coherence score. In the sentiment analysis, four different models are built, which their performance is assessed on in the evaluation.

The proposed system is evaluated based on the performance of the models. This evaluation is carried out through a number of experiments. In the first experiment, a baseline is determined for each classifier using two different baseline methods. These baseline results are compared with each other. This comparison is between the two different strategies but also between the different datasets. The highest performance of the baseline model is the lower limit when comparing the more complex models. In the 2nd experiment, the performances of the classifiers; Multinomial Logistic Regression, Naive Bayes, Decision Trees and Random Forest; are compared using the evaluation metrics. This experiment is performed for two different datasets, which are datasets that have also been used for research papers. In the last step the performances of the classifiers; Multinomial Logistic Regression, Naive Bayes, Decision Trees and Random Forest; are compared with the performances of the same classifiers in an already existing research paper. A research paper that has built a similar system. The final model is chosen based on the results of these experiments. All these experiments point to the fact that this system strives to obtain the best performance. A performance that is competitive with system performances from the literature.

6.2 Answers found for the research questions

The answers to the research questions and sub questions can be found in the results which are given in Chapter 5. The first experiment result answers the second sub question: "To what extent do company policies influence the values as user satisfaction, affordability and willingness for wireless mobile services? ". Figure 8 shows that Infinity Mobile has a much more negative sentiment for all values than Mint Mobile. For the value willingness, Mint Mobile even has a neutral sentiment, while Infinity Mobile has a negative sentiment. The difference in policies between the companies is because Infinity mobile opts for the traditional fixed plan approach[25], while Mint Mobile opts for more flexible plans for their customers[24].

So, consumers seem to have less negative sentiment for Mint Mobile's more flexible approach to providing services. This leads to a less negative sentiment for user satisfaction, affordability and willingness. It is clear from this that a business strategy does influence values such as user satisfaction, affordability and willingness. However, the average values remain negative or just about neutral, how much does this have to do with the current pandemic? This is the second sub question investigated. The time analysis also called the third experiment result provides an answer to this question. This analysis has shown that the sentiment for the value user satisfaction has hardly changed as a result of the pandemic. The variance in the sentiment values is small and therefore remains negative. With regard to affordability, there is clearly a trend break at the start of the pandemic. Before the pandemic, negative sentiment appeared to be moving towards neutral, however, the pandemic has turned this trend

into a downtrend. The sentiment for willingness was positive before the pandemic and negative after the start of the pandemic. However, it can be clearly seen that pre-pandemic sentiment was already trending downward towards neutral sentiment. Around the start of the COVID pandemic, a clear trend break can be observed. The downward trend just before the pandemic initially appeared to be decreasing, but at the start of the pandemic, a downward trend is set again towards a sentiment value of -1. Finally, the main research question was answered by the second experiment result of the evaluation. Figure 10 shows the ratio of negative, neutral and positive sentiment for each value. By looking at the tweets of the months with exceptionally high or low sentiment scores per value, the following things have been observed. As can be seen from figure 10 the sentiment for user satisfaction is mainly negative. Despite the fact that there are enough positive statements from consumers about the positive influence of wireless mobile services on telephone accessibility and therefore safety, there are many negative statements. These statements are mainly in the service domain, consumers criticize the poor customer service, the poor internet quality but also the lack of flexibility of the services.

For affordability, consumers are positive about the flexibility of the services, i.e. being able to choose which service you need. This can lead to a cost reduction. However, the pandemic still plays an important role for this value, from the statements it can be seen that people complain about the price because their own financial situation has deteriorated. Consumers also complain about the fact that wireless mobile providers do not take the financial situation of the customers into account. The willingness to use or switch from a service provider has also decreased due to the pandemic. Consumers are making hesitant statements about switching due to the uncertainty of the pandemic. Before the pandemic, consumers regularly indicated that they would switch their respective provider to their competitor if the service did not improve. Many comparisons were also made with companies such as Verizon or Virgin.

6.3 Limitations

There are some limitations for this research. The first limitation is that the social media platform Twitter was used to collect the data. It's assumed that the users of Twitter are a representative target group for the users of wireless mobile services. Research by Statistics Netherlands[9] has shown that in reality young people are over-represented on this social media platform. However, this research will be an indication of the actual sentiment.

Another limitation is that only English tweets could be examined. This allows you to only measure the sentiment of English-speaking users. Non-English tweets would cause problems in the sentiment analysis as the classifier would label these tweets as neutral. For topic modelling, these tweets would pose no problems. In addition, there are also a number of limitations to the system architecture. The result of this system is highly dependent on the quality of the data. The quality of the data is determined in the data extraction, cleaning and preparation steps. In the data preparation step, the dimensionality reduction can damage the quality of the data by removing valuable information from the dataset before the analysis. Another limitation is that the functioning of a topic model is highly dependent on the number of iterations and the number of topics chosen. Too few topics can result in multiple topics in one topic.

This can also lead to a distorted end result of the system.

6.4 Ethical considerations

There are a number of ethical considerations that must be taken into account during the research. After collecting the data, there were attributes containing personal information about the users. This concerns the username and the tweet attributes. From the username it is usually easy to find out what the real name of the user is. This is sensitive information that should be handled with care. The same goes for the tweet attribute. Twitter users regularly give away personal information in their tweets without realizing that this information can be used for wrong purposes. That is why the privacy of the users was treated very carefully in this study. The anonymity of the users is guaranteed and the data is only used for this research and is not shared with third parties. These steps meet the requirements of the GDPR.

In addition, ethical considerations must also be taken into account during the classification process, since it concerns information about real individuals. Decisions or statements can be made based on the classification; these statements must also be ethically responsible. This can be achieved by not including privacy sensitive data such as username/name in the classification process. That's why I removed this column before the classification. Tweets officially do not belong to sensitive attributes according to the law, so they can be included in the classification process. However, the rules regarding anonymity and privacy continue to apply. Lastly, work by other researchers I have used for this study, has been referenced through APA.

6.5 Further work

The research can be expanded or modified in a number of ways. In this research, the data was collected from one platform/medium called Twitter. As mentioned before, the system is very sensitive to the quality of the data. In further research it is possible to mine the data on this subject from different platforms and to compare the functioning of the system with each other. The evaluation can also examine which models work best for the relevant platform. Another extension of the research is to obtain data such as the gender of the users, the location of the users and the age of the users. This data must be handled very carefully as these are sensitive attributes and the analyzes must comply with the rules of the GDPR. For example, the location data makes it possible to compare the Western world with the Asian world. This allows you to gain insight into the sentiment on a separate continent for user satisfaction, affordability and willingness. This analysis can also be performed for the gender data, whereby the sentiment of women and men for these values can be examined. This analysis can also be performed for the age data. Finally, it would also be possible to perform a combination of these analyses. For example, you could compare the sentiment of women under a certain age limit in the Western world with women under the same age limit in the Asian world. This would provide a better overview of the sentiment for the various values and make it more transparent.

Bibliography

[1] Manal Abdulaziz, Alanoud Alotaibi, Mashail Alsolamy, and Abeer Alabbas. "Topic based Sentiment Analysis for COVID-19 Tweets". In: *International Journal of Advanced Computer Science and Applications* 12.1 (2021).

[2] Nikolaos Aletras and Mark Stevenson. "Evaluating Topic Coherence Using Distributional Semantics". In: 2013, pages 13–22.

[3] Rubayyi Alghamdi and Khalid Alfalqi. "A Survey of Topic Modeling in Text Mining". In: *International Journal of Advanced Computer Science and Applications* 6 (2015).

[4] Jehad Ali, Rehanullah Khan, Nasir Ahmad, and Imran Maqsood. "Random Forests and Decision Trees". In: *International Journal of Computer Science Issues(IJCSI)* 9 (2012).

[5] Murugan Anandarajan, Chelsey Hill, and Thomas Nolan. "Semantic Space Representation and Latent Semantic Analysis: Maximizing the Value of Text Data". In: 2019, pages 77–91.

[6] Erkan Arı. "Using Multinomial Logistic Regression to Examine the Relationship Between Children's Work Status and Demographic Characteristics". In: *Research Journal of Politics, Economics and Management* 4 (2016), pages 77–93.

[7] Ronald Beaubrun and Samuel Pierre. "Technological developments and socio-economic issues of wireless mobile communications". In: *Telematics and Informatics* 18 (2001), pages 143–158.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *J. Mach. Learn. Res.* 3 (2003), pages 993–1022.

[9] CBS. "ICT,kennis en economie". In: *CBS* 4 (2018).

[10] Emerce. *Negentig procent van de wereldbevolking heeft mobiel in 2020.* `https://www.emerce.nl/nieuws/negentig-procent-wereldbevolking-heeft-mobiel-2020`. Nov. 2014.

[11] Peter Foltz. "Latent Semantic Analysis for Text-Based Research". In: *Behavior Research Methods* 28 (1996), pages 197–202.

[12] Giri. *Topic model evaluation.* `https://highdemandskills.com/topic-model-evaluation/`. June 2021.

[13] Matthew Hoffman, David Blei, and Perry Cook. "Content-Based Musical Similarity Computation using the Hierarchical Dirichlet Process." In: *ISMIR 2008 - 9th International Conference on Music Information Retrieval* (2008), pages 349–354.

[14] Chanmi Hong and Lisa Slevitch. "Determinants of Customer Satisfaction and Willingness to Use Self-Service Kiosks in the Hotel Industry". In: *Journal of Tourism & Hospitality* 07 (2018).

[15] Mohammad Hossin and Sulaiman M.N. “A Review on Evaluation Metrics for Data Classification Evaluations”. In: *International Journal of Data Mining & Knowledge Management Process* 5 (2015), pages 01–11.

[16] H. Jelodar, Yongli Wang, Chi Yuan, and Xia Feng. “Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey”. In: *Multimedia Tools and Applications* 78 (2018), pages 15169–15211.

[17] Chhinder Kaur and Anand Sharma. “Twitter Sentiment Analysis on Coronavirus using Textblob”. In: (2020).

[18] Hidetoshi Kawaguchi. “Dummy training data generation method towards robust estimation of confidence value of semi- automatic agents for multi-class classification”. In: (2019).

[19] Brian Keith Norambuena, Exequiel Lettura, and Claudio Villegas. “Sentiment analysis and opinion mining applied to scientific paper reviews”. In: *Intelligent Data Analysis* 23 (2019), pages 191–214.

[20] Yu-Cheng Lee, Yu-Che Wang, Shu-Chiung Lu, Yi-Fang Hsieh, Chih-Hung Chien, Sang-Bing Tsai, and Weiwei Dong. “An empirical research on customer satisfaction study: a consideration of different levels of performance”. In: *SpringerPlus* 5 (2016).

[21] Yang Liu, Jian-Wu Bi, and Zhi-Ping Fan. “Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms”. In: *Expert Systems with Applications* 80 (2021).

[22] D.I. Martin and M.W. Berry. “Mathematical Foundations behind Latent Semantic Analysis”. In: 2007, pages 35–55.

[23] Kaiz Merchant and Yash Pande. “NLP Based Latent Semantic Analysis for Legal Text Summarization”. In: *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. 2018, pages 1803–1807.

[24] Mint Mobile. *How it works*. `https://www.mintmobile.com/plans/`. 2021.

[25] Xfinity Mobile. *About us*. `https://www.xfinity.com/mobile/`. 2021.

[26] Edi Surya Negara and Dendi Triadi. “Topic Modelling Twitter Data with Latent Dirichlet Allocation Method”. In: 2019, pages 386–390.

[27] Subramani Parasuraman, Sam Thamby, Yee SW, Chuon BL, and Lee Yu Ren. “Smartphone usage and increased risk of mobile phone addiction: A concurrent study”. In: *International Journal of Pharmaceutical Investigation* 7 (2017), pages 125–31.

[28] Rajul Parikh, Annie Mathai, Shefali Parikh, Ganja Sekhar, and Thomas Ravi. “Understanding and using sensitivity, specificity and predictive values”. In: *Indian journal of ophthalmology* 56 (2008).

[29] Hitesh Parmar, Sanjay Bhanderi, and Glory Shah. “Sentiment Mining of Movie Reviews using Random Forest with Tuned Hyperparameters”. In: 2014.

[30] Louis-Francois Pau. “Mobile Service Affordability for the Needy, Addiction, and ICT Policy Implications”. In: 2008, pages 41–48.

[31] Ankita Rane and Anand Kumar. “Sentiment Classification System of Twitter Data for US Airline Service Analysis”. In: *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*. Volume 01. 2018, pages 769–773.

[32] Irina Rish. “An Empirical Study of the Naïve Bayes Classifier”. In: *IJCAI 2001 Work Empir Methods Artif Intell* 3 (Jan. 2001).

[33] Muhammad Sarwar and Tariq Soomro. “Impact of Smartphone’s on Society”. In: *European Journal of Scientific Research* 98 (2013).

[34] Arun K Sharma. *Multinomial Logistic Regression.* https://www.mygreatlearning.com/blog/multinomial-logistic-regression/. 2021.

[35] Arun K Sharma. *Understanding Latent Dirichlet Allocation.* https://www.mygreatlearning.com/blog/understanding-latent-dirichlet-allocation/. 2020.

[36] Keng Siau and Zixing Shen. "Mobile Communications and Mobile Services". In: *Int. J. Mob. Commun.* 1 (2003), pages 3–14.

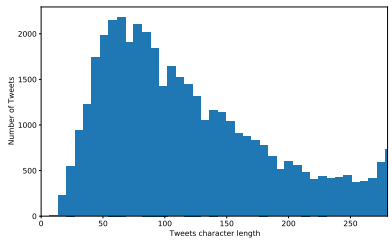
[37] Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. "Exploring Topic Coherence over many models and many topics". In: July 2012.

[38] Yee Teh, Michael Jordan, Matthew Beal, and David Blei. "Hierarchical Dirichlet Processes". In: *Machine Learning* (2006), pages 1–30.

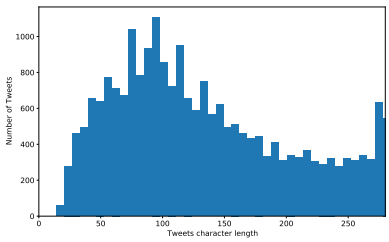
Appendix A

Appendix: More Figures.

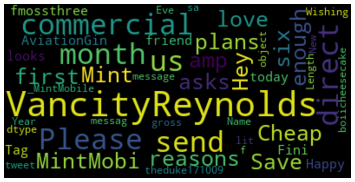
A.1 Appendix A



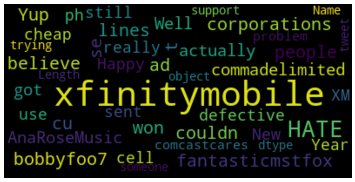
Distribution tweet length Mint Mobile



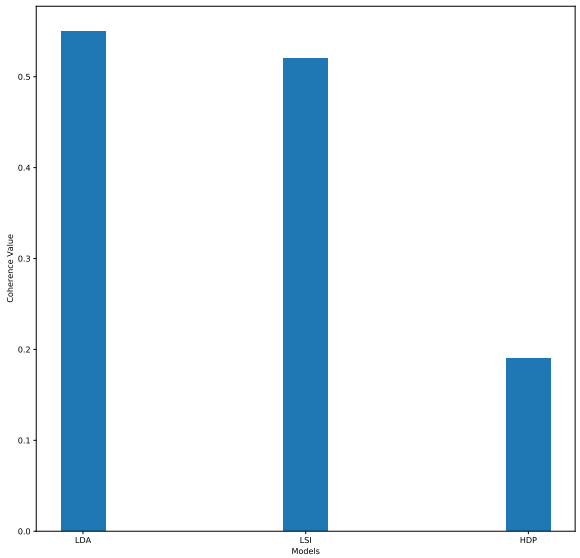
Distribution tweet length Xfinity Mobile



Word cloud Mint Mobile



Word cloud Xfinity Mobile

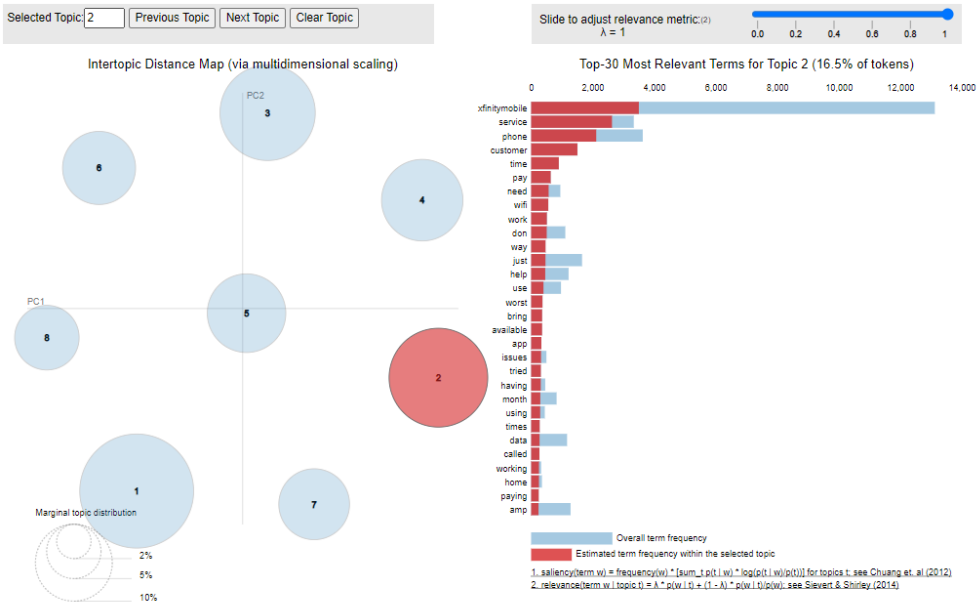


Bar plot coherence scores

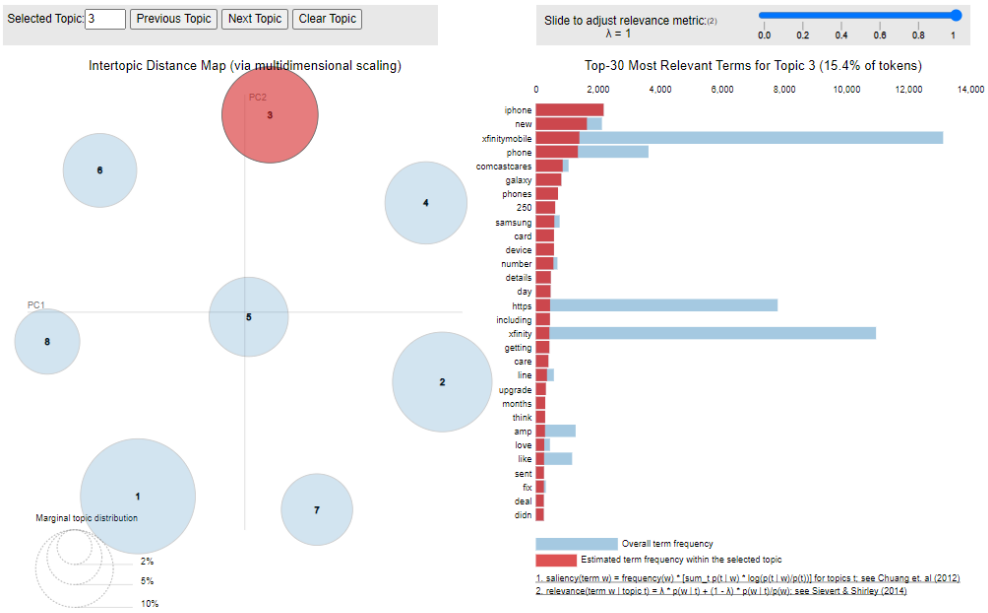
Topic models	Coherence scores
LDA	0.55
LSI	0.52
HDP	0.19

Coherence score topic models

A.2 Appendix B



Topic modelling visualization affordability for Mint Mobile



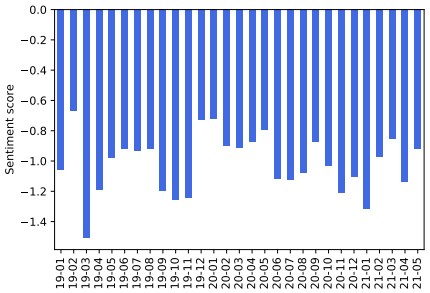
Topic modelling visualization willingness for Mint Mobile

A.3 Appendix C

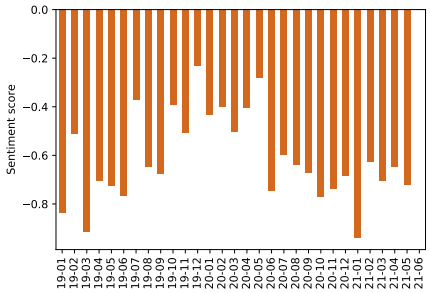
User satisfaction Tweets
'Still waiting for mintmobile to honor their 7-day money back guarantee from 01/14/21. I 1st tweeted mid 04/21, now 05/31/21. There is no excuse to take over a month when I have provided my receipt and information here and through emails to mint mobile. Mismanag @VancityReynolds'
'@dbrower @Mintmobile Mint buys wholesale extra network capacity access from T-Mobile at wholesale prices and resells it. Mint is a middle-man making a deal for extra data capacity T-Mobile would be wasting if someone else didn't make a deal for it.'
'@Mintmobile The 5gb hotspot cap on on the unlimited plan was a deal breaker for me.'
'@VancityReynolds Wish I could feel as good about my #MintMobile The customer service is horrible... spent hours with multiple people and all they offered was "We're sorry you aren't pleased, he's a 10 dollar credit (Note: ONLY IF YOU RENEW SERVICE) That's one way to try and keep us, NOT! #RyanReynolds'
'@Mintmobile lovely service!!'

Affordability Tweets
'Looking to save some serious cash on your cell phone? Give @MintMobile a try! It's great, uses TMobile network and saves tons of #money! #savings #cellular #savebig #rt'
'@dknycbd @JohnLegere @TMobile @Mintmobile They have a plan for 55+ with Netflix on us is very cheap'
'@SixteenthLight @Verizon @Mintmobile If you want to save 15 dollars more on your first renewal (not the initial plan purchase), you can use the referral link in my twitter profile. They're also running a 3 months free promo right now, so yeah its 45+ dollar tax for 6 months. Double check the coverage map in my profile. ty'
'Just switched to @MintMobile ,No dead spot in my home. Still need to check my office but already smiling with the budged low price. Hehe '
"Switch to @Mintmobile and save lots of dollars'

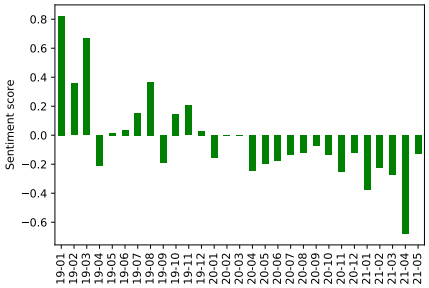
Willingness Tweets
'@MintMobile Any particular reason why my bill has slowly shot up by $10 since last year in the checkout's hidden fees department? I've been with you guys since 2017, but if it climbs any higher and I'll go somewhere else.'
'@VancityReynolds just signed up for @mintmobile. If it's sucks; I'm sending @deadpool after you!'
'OK I would probably switch to @Mintmobile if I could just because of @VancityReynolds, this ad, and more importantly, this desktop. Fav folders: SadRyan/BlueSteel/SmartRyan and Books to say I've read. '
'"@darklands23 We're so glad you made the switch! '
"Officially switching to @Mintmobile! The fact that @VancityReynolds (or Berg as my wife and I refer to him) is an owner caught my attention, but the commercial with my parenting role model, Rick Moranis sealed the deal. All the money we will be saving helped a bit too.'



Monthly mean sentiment user satisfaction



Monthly mean sentiment affordability



Monthly mean sentiment Willingness

A.4 Appendix D

Code can be found in the GitHub link:
https://github.com/qahtanaa/SATAF_Tweets