Master Dissertation Proposal

Social Network Analysis via a Sentiment Dictionary for identifying the most promising service opportunities within a local and industry-specific context: The case for small to medium-sized hotels

Table of Contents

0.) Abstract

Through the literature review conducted in this academic article at least one essential gap is discovered. Namely the absence of extracting customer insights via social media analytics by small to medium-sized hotels (SMSH) in order to enhance their services. Therefore the proposed research aims to demonstrate how an inexpensive and feasible framework may obtain social media reviews and it outlines the systematic design of establishing sentiment classification capability for the collected text data in order to output not just the services where improvement is required, but recommendations of how to refine a particular service of interest, while considering the resource limitations of SMSH. For that purpose the creation of a local sentiment dictionary will be suggested, the utilisation of the Naive Bayes classifier is going to be advised, the implementation of the rapid automatic keyword extraction algorithm will be proposed and last but not least LDA topic modelling is advocated due to its big data suitability. In addition the Dash library is regarded as beneficial for data visualisation addressed towards SMSH due to its interactive and modern characteristics. Moreover the pertinence of Yelp data for training and testing is discussed, while indicating that actual Twitter data presents the most appropriate repository for gathering negatively connoted customer data. Finally the master dissertation schedule is explained and illustrated.

1.) Introduction

From the commencement of operations by social networks until today, social media (SM) has attracted an extensive amount of users, who are sharing each day their thoughts in form of visual, textual or audio data. Examinations of this trend by Koiranen et al. (2019) were illustrating that between 2008 and 2016 the likelihood to own a SM account rose across all socio-demographic milieus in Finland, which proves to be consistent with the results by Ofcom (2017), who could determine an elevated exposure to social networks for internet users of every age in the UK for the timespan of 2012 to 2016. Despite the increase in adoption rates, Koiranen et al. (2019) outline that the growth was non-symmetrical among all age groups due to the relatively accelerated acceptance proportions of the younger brackets in comparison to the elderly. In addition Ofcom (2017) was portraying that unambiguous social grade classes exhibit a custom taste for the variety of SM platforms, when being set in contrast to the remaining social grade brackets, respectively.

According to Lee (2018) the application purpose of social media analytics (SMA) is defined as follows: “Social media analytics are used to monitor and listen to word-of-mouth that spreads in social media platforms, and conduct thorough analyses of consumer opinions on products and services“. While small to medium-sized enterprises (SMEs) on the one hand reveal a demand for customer insights for the enhancement of their products or services or both ((Ahmad, Ahmad and Bakar, 2018), (Liu et al., 2020)) under the prerequisite of meeting exceptional data privacy and data security standards (Rajabion, 2018), on the other hand they are confronted with the allocation of their strictly limited resources in terms of budget expenses (Ahmad, Ahmad and Bakar, 2018), technological know-how and human capital (Liu et al., 2020). Furthermore SMEs are categorised according to Rajabion (2018) in the succeeding fashion: “SMEs in the United States refers to organisations with employees that are fewer than 500“. Moreover the relevant undersupply of professionals being capable of conducting SMA and analysing its results in the USA (Rajabion, 2018) as well as the significance of sentiment analysis research for the tourism industry (Kirilenko et al., 2018) highlights the essentiality for feasible SMA approaches in the hospitality sector especially for small to medium-sized hotels.

2.) Relevance and design of the social media analytics master dissertation

2.1.) Literature review

Ulwick (2005) created for companies the opportunity formula, which is offering enterprises the assistance to prioritise distinct business opportunities for which the maximum level of importance and the least degree of satisfaction is required in order to be ranked as the most relevant opportunity (as can be withdrawn from Formula 1):

$$Opportunity\,Score = Importance + Max(Importance - Satisfaction, 0)$$

Formula 1:
The opportunity formula by Ulwick (2005)

Furthermore Hu et al. (2017) discovered in their study that positive sentiment was mainly directed towards the producing sector of an economy, whereas in sharp contrast the service sector received negative sentiment to a massive extend, which is consistent with the observations made by Farizah Ibrahim and Wang (2019), that indicate immense negativity for delivery and customer services in the retail industry. Despite these potentials the researchers Jeong and Yoon and Lee (2019) applied their opportunity mining framework (which bases on the opportunity algorithm by Ulwick (2005)) exclusively on a single multi-functional product while missing the approach's suitability for services due to the greater degree of dissatisfaction therefore opening up more room for improvements although Jeong and Yoon and Lee (2019) did not perform their method on hybrid product-service systems as well. In fact Ahmad, Ahmad and Bakar (2018) were detecting, that SMEs are in need for listing to customers' voices in order to create novel products and service, to evaluate current services and items as well as to broaden their knowledge basis for taking well-informed decisions. Equally Liu et al. (2020) were confirming that SMEs were eager to execute big data analytics for customer preference extraction. To illustrate even further the researchers Liu and Burns and Hou (2017) generally advocated for advanced customer service development driven by SMA, while Davras and Caber (2019) were strongly requesting such a model in order to quantify the individual impact of particular service features on customer satisfaction especially for the hotel industry whereas Farizah Ibrahim and Wang (2019) identified tremendous opportunities for SMA limited to the retail industry, which is not including the hospitality industry as a retail service according to the Office of Management and Budget (2017).

Although big data analytics is only exercised by circa 50% of the US SMEs ((Statista Research Department, 2016), (Deloitte, 2019)), implementing these evaluation techniques benefits SMEs in terms of finance as well as marketing (Maroufkhani et al., 2020). In addition Lee (2018) states that businesses in general may elevate their relative market position by conducting SMA through capturing a more detailed picture of competitors and their respective customers as well as applying the same to their own customers resulting in a information richer foundation for the analysis of the business environment. Furthermore Dong and Yang (2020) brought to light that SMEs profit more in terms of market performance than larger firms from an adoption of big data analytics if they are also exhibiting a presence on a wide range of social media networks. Despite barriers of SMEs to

big data analysis integration like for instance confusion as well as uncertainty (Maroufkhani et al., 2020) and data security (Rajabion, 2018), the absence of proper data analysis in the domain of SMEs is prevalent for even small data that these enterprises possess according to Liu et al. (2020) leading to a potentially depressed refinement of products and services (Liu et al., 2020). While Liu et al. (2020) is recommending for this reason a cloud-based architecture, one has to be mindful that trust and dependency in a third-party provider like Amazon Web Services (AWS) might first disappoint since Amazon was creating its own music streaming service after Spotify choose to host their service on Amazon's infrastructure, which is also true for movie and tv-shows for the company Netflix, and secondly making the SMEs business more fragile to be shut down by third-party cloud providers like the social media network Parler experienced with AWS, so that only a SME owned deployment seems reasonable, if a SME is looking for a cloud architecture solution.

When determining the most appropriate social network for collecting service data Blagus and Zitnik (2018) recorded that while Twitter contained the most negative texts among the five monitored social networks, the ratio of text which could be assigned to a sentiment value was relatively low with 24%, which is close to the 17% of non-neutral tweets identified by Farizah Ibrahim and Wang (2019). Controversially Liu and Burns and Hou (2017) was able to link 66% of their 1.7 million tweets to sentiments, while 47.8% of their entire dataset could be associated with negativity, which was approximately thrice as much as the occurrence rate of positivity (Liu and Burns and Hou, 2017). In fact Blagus and Zitnik (2018) admitted that Twitter was still the most impactful social media among the five. Finally these findings are supplemented by Singh and Verma (2020), who were concluding that tweet's author metadata delivered no significant improvement as well as that tweets without textual data are irrelevant for SMA, whereas on the other hand Molinillo et al. (2019) assert that the most frequent way of communication in social networks tends to be more passive (i.e. likes and shares) than active (i.e. comments).

2.2.) Aims and Objectives

Twitter contains an overabundance of customer opinions, which express negative feedback to an extensive degree ((Liu and Burns and Hou, 2017), (Blagus and Zitnik, 2018)) towards services (Hu et al., 2017) leading to the potential beneficial discovery of new and customisation of existing customer services ((Lee, 2018), (Ulwick, 2005), (Liu and Burns and Hou, 2017)) and specific analysis for each service feature in the domain of hospitality (Davras and Caber, 2019). Neglecting to analyse these valuable perspectives like approximately 50% of US SMEs ((Statista Research Department, 2016), (Deloitte, 2019)) leads to competitive disadvantages (Lee, 2018) for small to medium-sized businesses (Maroufkhani et al., 2020), although their strict resource limitations ((Ahmad, Ahmad and Bakar, 2018), (Liu et al., 2020)), technological know-how ((Liu et al., 2020), (Maroufkhani et al., 2020)) and data security standards (Rajabion, 2018) have to be considered. Thus the following research question arises: May the opportunity algorithm be deployed for multi-functional services under real-world conditions from a small to medium-sized hotel's point of view, thereby firstly detecting important and unsatisfied services and secondly indicating improvement directions for individual service features? To answer the aforementioned question this research aims to present an affordable as well as effective architecture for capturing social media texts and for evaluating their corresponding sentiments in order to generate decision support for service optimisation in the hotel industry. This project is going to create (under the assistance of performing a machine learning algorithm on Yelp data) a domain-specific sentiment dictionary, implement the RAKE keyword extraction algorithm on Twitter data in order to conduct latent dirichlet topic modelling (LDA) for calculating the opportunity scores of services thereby indicating enhancement recommendations for each particular service.

This aim is further specified into objectives (as depicted in Graphic 1), which are sorted in a chronological order so that the first objective, is the first one to be achieved before proceeding to the next one. Therefore first the Yelp dataset is required to be obtained. The key reason for selecting the Yelp dataset for training the machine learning classifier over Twitter data is that, the Yelp dataset comes with already customer-sentiment labelled hotel reviews, whereas Twitter data has to be categorised by the researcher into sentiment classes manually or automatically for instance via emoticons programmed by the academic, which is highly likely to be inferior to the ratings by customers themselves on Yelp. In addition the Yelp dataset is able to be filtered in order to be domain-specific in terms of location (city) and in terms of business categories like the hotel sector. While Davras and Caber (2019) reported eight distinct service dimensions, when inspecting a hotel in Antalya (Turkey), Nakayama and Wan (2019) emphasised the point of a locally confined data analysis through their findings, in which they portrayed that ethnic culture influences the evaluation of customer reviews, resulting in a competitive advantage for domain-specific sentiment classifiers in comparison to general sentiment classifiers, which were also applied in the opportunity architecture by Jeong and Yoon and Lee (2019). In fact according to Kirilenko et al. (2018) it is optimal to train machine learning algorithms as well as sentiment lexicons on domain-specific data due to the boosted validity to adjust to the custom input by users (Li and Li and Jin, 2020), which may culminate in more sophisticated services (Li and Li and Jin, 2020).

Graphic 1:
Aim, Objectives and Methods for the Master Dissertation

Next comes the training of the supervised Naive Bayes algorithm and it's performance evaluation in order to insure that the Naive Bayes classifier satisfies at least the condition of acceptable classification standards on the proxy Yelp test dataset, which comprises the characteristic of holding the correct sentiment rating by the customers attached to the reviewed text and is therefore assumed to be representative for the Twitter data collected latter in the process. Last but not least if the quality of Naive Bayes suffices minimum classification standards, a sentiment dictionary is constructed holding the sentiment value for each word, which may be assigned to one of the following classes: Extremely unsatisfied (denoted as 1), slightly unsatisfied (denoted as 2), neutral (denoted as 3), slightly satisfied (denoted as 4) or extremely satisfied (denoted as 5).

Having accomplished the sentiment dictionary generation, historical tweets from 2019 for the hotel industry is collected by searching domain-specific hashtags. The reasoning behind favouring the year 2019 over more recent data from 2020 is to avoid the externalities from the year 2020, namely the outbreak and spread of Covid-19 and the exceptional government responses for the hospitality sector. The upcoming objective (as illustrated in Graphic 1) is extracting the keywords from each individual tweet, while the succeeding objective concerns proper data cleaning, which is essential for assigning keywords to the LDA topics and for the correct manual labelling of each topic. One further objective is to compute the number of topics so that the semantic similarity among the keywords for one topic should be as homogeneous as possible whereas the semantic similarity among the various LDA topics would be ideally as heterogenous as thinkable. Finally an adequate topic modelling technique is performed, which is also scalable to deal with big data in terms of its volume. Therefore this subgoal builds the basis for first calculating the importance of a topic and for second constructing a sentiment-keywords to topics matrix, which are the key input factors for the calculation of the opportunity scores for each topic. Thereafter the opportunity scores are required to be evaluated, so that the highest opportunity value represents the most likely service area for which refinement by the company is the most rewarded by the customer opinions. In addition one objective should lie in inspecting the topic's keywords for which the one with the most negative connotation serves as recommendation for the enhancement of this specific topic field (Jeong and Yoon and Lee, 2019). The final objective presents a more attractable fashion for the human eye to comprehend project data in shape of graphics and even interactive graphics, so that complexity reduction is accomplished.

Three major limitations of these approach have been set. First the execution of the presented algorithm is not intended to be performed in a scalable setup in its current form since the author of this master dissertation proposal planned to imitate SME's conditions, so under the strictest financial and technological requirements for conducting SMA is going to be operated. Second according to the criteria by Mukhtar and Abid Khan and Chiragh (2018) the proposed sentiment lexicon is not able to handle negation of a word for example "did not like" nor amplifying words like for instance "hate very much", however one could argue that Yelp customers were assigning their sentiment to their own texts, therefore the words of review should exhibit a tendency towards the proclaimed review sentiment in general. Another criterion by Mukhtar and Abid Khan and Chiragh (2018) is addressing the semantic meaning of a word based on its context. This is indeed the Achilles' heel of of the Naive Bayes algorithm, which assumes that each word in a sentence is

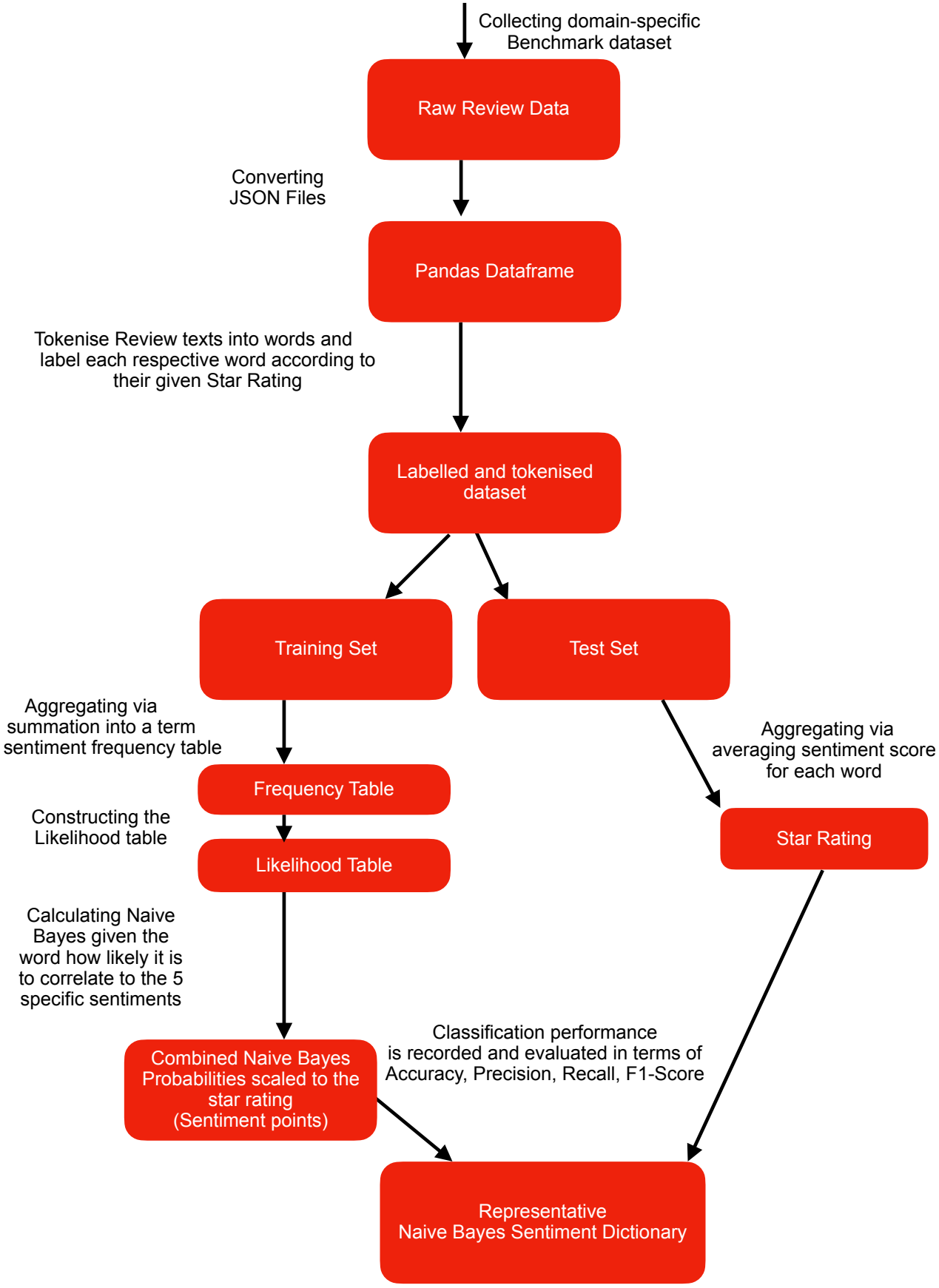
generated absolutely independent of any prior or subsequent word, thereby neglecting the context of a word in the realm of natural language processing (NLP). Third the suggested technique does not differentiate between novel customers and regulars unlike the procedure realised by Davras and Caber (2019).

2.3.) Methodology

The section will discuss the fashion and especially the reasoning behind the implementation of a particular approach by proceeding in the chronological order of the opportunity process, which is pictured in total in Graphic 3. The originator of this academic paper will therefore commence the discussion about the usage of APIs (Application Programming Interfaces) for data collection.

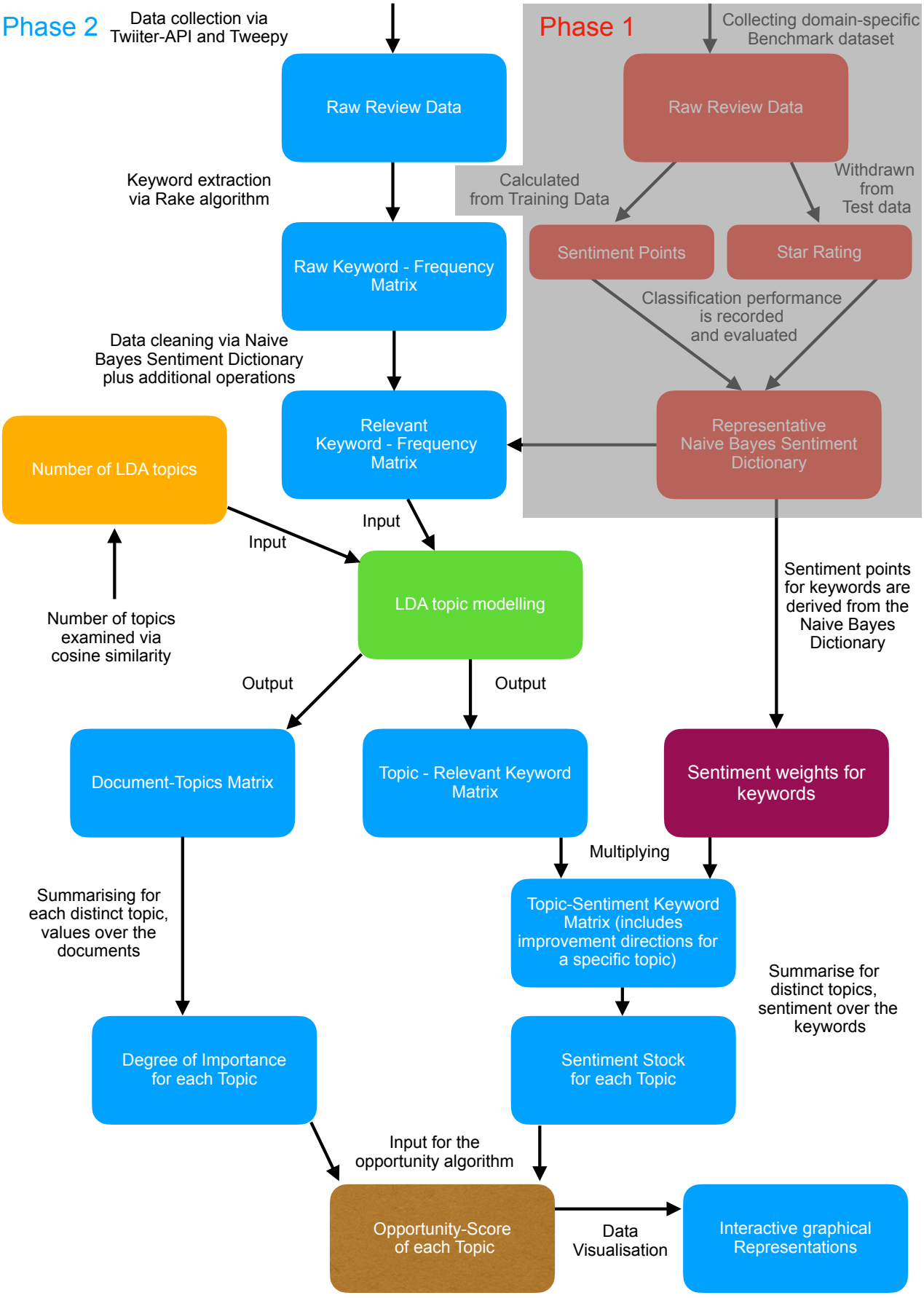
The rationale for selecting the Yelp dataset only for training and testing the sentiment classifier and thereby creating the sentiment dictionary in “Phase 1” (as may be extracted from Graphic 2) is that first as mentioned above its characteristic of containing already sentiment-labelled scores by actual customers linked to their respective customer reviews whereas Twitter’s text data is missing these sentiment classifications although these may be assigned manually, which is time-consuming and may also not reflect the true intention of the tweet’s originator at all times, or these may be automatically executed via allocating sentiment based on emoticons like it is conducted by Li and Fleyeh (2018) as well as Li and Li and Jin (2020), which is as well unlikely to perform a perfect sentiment classification for the entirety of customer reviews like only the review writers are capable of. The second point concerning why Yelp data is not collected as input in “Phase 2” (like portrayed in Graphic 3) instead of Twitter data is explained through the strict limitations of the Yelp Fusion API, which just emits three customer reviews per request (Yelp, 2021), while Twitter transmits at the maximum between 100 to 500 tweets per request (Twitter, 2021b). Furthermore Twitter offers a premium search archive function, which is also available for enterprises as of the writing date of this report (27th of January 2021) according to Twitter (2021a). Moreover this premium search archive function presents a feasible way of accessing past Twitter data (Twitter, 2021a) of hotels, which were formulated in a relatively regular environment i.e. before the outbreak of the Covid-19 crisis and thereby prior to the start of unprecedented government directives towards the hospitality sector.

Next when examining whether to operate this approach in the programming language R or in Python according to the recommendations by Gallagher and Trendafilov (2018) it is more suitable to employ Python since both programming languages are sophisticated in order to analyse big data, however as most SMEs are going to be beginners in SMA, the simplicity of Python is categorised as more ideal than the one of the programming language R (Gallagher and Trendafilov, 2018). Therefore the Python pandas and numpy libraries are going to be called in order to read, convert and select the data for “Phase 1” into a pandas dataframe. In the following the NLTK package will be utilised in order to tokenise the individual tweets into words as well as for stop words and punctuation removal and the pandas dataframe explode function will assign the specific sentiment to each word, which is likely to lead to multiple repetitions, because words like “love” or “hate” are probably going to be used in various sentences of the documents. Therefore it is required that these different sentiments for a word are aggregated latter for the training and test set respectively. Furthermore the “Phase 1” dataset is going to be divided into training and test set, while the training set is constructed to be representative for the “Phase 1” dataset in terms of sentiment distribution by conducting proportional sampling, which is defined according to Etikan and Bala (2017) as follows: “Proportional allocation is used when the sample size from different stratum will be kept proportional to the strata size”. In addition Barragán-Landy et al. (2020) state



Graphic 2:

The sentiment dictionary creation (Phase 1)



Graphic 3:

The entire opportunity process (including the sentiment dictionary) on SM data

that this method maintains the properties of the original dataset so that the chosen sample may be classified as representative.

The Naive Bayes classifier will in the following step examine the probabilities for a given word that the certain review belongs to the five sentiment classes mentioned in the last section. Why Naive Bayes was selected as a classifier is inferred by the assed literature. Elzayady, Badran and Salama (2018) were outlining that the methods Naive Bayes and logistic regression were surpassing decision trees respectively, which was affirmed by Kunal et al. (2018), who were deducing that Naive Bayes reached an enhanced precision compared to the method of decision trees plus that the merits of Naive Bayes of being topic as well as language-independent were emphasised, although it is required to adjust the list of stop words correspondingly to the examined language. Kritiyanti et al. (2018) were deducing that Naive Bayes performed to a higher satisfying degree compared to support vector machines (SVM) the job of categorising public sentiment towards political candidates under scarce data conditions, whereas Kirilenko et al. (2018) found that Naive Bayes and SVM respectively are executing classification tasks well, if the training data, which serves as input for both algorithms, is belonging to the same domain as the testing data.

Krouska and Troussas and Virvou (2017) discovered that Naive Bayes and SVM are superior in analysing the sentiment of tweets than the methods K-nearest neighbours (KNN), logistic regression, decision trees and the dictionary method SentiStrength, whereas Kirilenko et al. (2018) recommend sentiment dictionary approaches like SentiStrength if the implementation of machine learning algorithms is not an option due to financial or temporal restrictions. Furthermore Mukhtar and Abid Khan and Chiragh (2018) attained sentiment classification results, which outline that their lexicon method was more effective as well as economical than the machine learning techniques decision trees, KNN and SVM, although it was not being set into relation with Naive Bayes. Controversially Li and Li and Jin (2020) deduced from their experiments that first their Naive Bayes sentiment dictionary was surpassing N-gram Naive Bayes, SVM and textual convolutional neural networks especially for negative sentiment in terms of precision, accuracy, F1-score while attaining the second best result for recall, and second Li and Li and Jin (2020) regarded Naive Bayes as simple, but reliable for sentiment allocation undertakings. In contrast Jain and Kaushal (2018) were reasoning based on their examinations that deep learning algorithms delivered more preferable outcomes in terms of sentiment classification than machine learning algorithms (like also Naive Bayes), although Jain and Kaushal (2018) admitted that machine learning algorithms may be favoured, if the metrical classification differences towards a deep learning model are not outstanding, due to the lower cost of realisation for these type of approaches. In addition Rustam et al. (2019) brought to light that the Gaussian variant of Naive Bayes underperformed in terms of accuracy their remaining examined methods in conjunction with multiple keyword extraction algorithms, so that the Gaussian alternative was disregarded for this endeavour. Lastly Kirilenko et al. (2018) were revealing although machine learning and lexicon approaches are on a similar level to human raters for categorising noise absent datasets, they still lose accuracy if the datasets are becoming noisier, whereas Kirilenko et al. (2018) acknowledged that human raters in comparison to these classifiers possess also off dataset information, which is hard to prevent.

The Naive Bayes emitted sentiment probabilities for each word will be utilised as weight in order to calculate the total sentiment (i.e. sentiment point) for each word respectively. Therefore each sentiment point will be rounded in order to be clearly categorised into one sentiment class. Furthermore all sentiment points are going to be compared to their corresponding average star score of the same word in the test set. Then it will be evaluated via accuracy, precision, recall and F1-score, how well the classifier may estimate customer reviews. If the result is not sufficient more data cleaning is required in order to tweak the classifier. For the case that the outcome reaches minimum thresholds, the Naive Bayes sentiment points and their corresponding words will be saved in a key value structure, namely a dictionary.

After “Phase 1” is completed “Phase 2” as described in Graphic 3 is commencing with the collection of Twitter data with the help of hashtags. The next stage is including the data cleaning of Tweets identical to the following process by Rustam et al. (2019), which first involves the elimination of punctuation and numbers, second the lower casing, third the stemming and last but not least the removal of stop words. In addition results by Resyanto and Sibaroni and Romadhony (2019) have shown that lower casing and stemming are the most crucial moves in order to attain an enhance accuracy.

The motive behind implementing RAKE and not TF-IDF nor TextRank is that while precision plays a more relevant role for evaluating the Naive Bayes classifier, for keyword extraction recall is more significant. This can be explained the best if both possible errors are taken into account: The worst case for a precise classifier is for e.g. that more words contain actually negative sentiment (which is the searched class), than classified. On the other hand the worst case for a classifier with well recall results is for instance, that more words are classified as holding negative sentiment (which is the searched class) than there actually are. While a precise classifier leaves out actual negatively connoted words, while being sure those identified as negative are to a high percentage actually negative, the keyword extraction algorithm should aim to filter out a significant portion of the relevant words, even if some are irrelevant, which may be removed through latter data cleaning. According to the findings by Ganiger and Rajashekharaiah (2018), RAKE was selected due to its competitive advantage on the recall characteristic over the two remaining approaches. In addition the inventors of the opportunity process (Jeong and Yoon and Lee, 2019) were as well making use of this technique, which may lead to enhanced comparability. Furthermore Jeong and Yoon and Lee (2019) eliminated in their approach around 90% of RAKE candidate words and about 50% of their data was rejected for further analysis. Finally after the RAKE candidate keywords are produced (see Graphic 3), they will be filtered by the Naive Bayes sentiment dictionary in order to discard all words with a neutral sentiment or words of which no sentiment value is available in the lexicon.

Once a relevant-keyword matrix is created one may calculate the optimal number of topics via the elbow method (Wang et al., 2014), so that the cosine similarity between topics is as minimal as possible. This number is essential as an input factor for LDA topic modelling which succeeds shortly after. LDA topic modelling is a key component of this proposed architecture since it is a well-known machine learning technique, which is apt to deal with the vast amount of textual data (Liu and Burns and Hou, 2017). In fact Farizah Ibrahim and Wang (2019) were able through the

utilisation of this approach to identify two new topics, whereas the labelling of topics like it is conducted by Jeong and Yoon and Lee (2019) is coming with the drawback of human subjectivity. Last but not least Vayansky and Kumar (2020) provide a decision tree in order to guide through the various topic modelling models thereby confirming if an average number of words per document falls below 50 and multiplex topic connections are exhibited and keyword information is obtainable, that a LDA keyword summarisation approach is appropriate.

Next as is presented in Graphic 3 the LDA technique outputs the document-topics matrix and the topic-keyword matrix. Furthermore the document-topics matrix will be aggregated via addition of topic frequencies over the entirety of documents, whereas the topic-keyword matrix is first going to be multiplied by a vector consisting out of the sentiment point for each respective keyword. Moreover the entries of the topic-sentiment-keyword matrix will then be added over the keywords. After this action has been completed the topic-sentiment-keyword and the document-topics matrix will be transformed in order to range from 0 to 10, where 0 represents the least importance (extracted from the aggregated document-topics matrix) or the least satisfaction (calculated from the aggregated topic-sentiment-keyword matrix) and 10 determines the maximum for both dimensions respectively.

Once the opportunity scores are computed according to the Formula 1, they are able to being ranked, of which the top-scores will recommend the best opportunities in terms of satisfaction as well as importance and the topic-sentiment-keyword matrix will indicate the feature for which the least satisfaction is shown of a topic, so that it is beneficial to advance the development in this area. Last but not least the final data is required to be visualised in an attractive fashion in order that SMEs in the hotel industry may quickly withdraw which service(s) possesses the most opportunity and what feature(s) must be enhanced. According to Burch et al. (2019) the Python Dash library provides a simple user interface, while Singh and Verma (2020) are confirming that this library delivers state-of-the-art features like dropdown menus, sliders and graphs in a web browser. On the other hand the well-known Python Seaborn library will be utilised as well in order to create attractive documentation of the internal process and is therefore addressed towards non-users especially in the environment of small to medium-sized hotels.

2.4.) Project Plan

The schedule for the proposed master dissertation is portrayed in Graphic 4. In this timetable it is intended that the data collection and the corresponding formatting of Yelp data (14th of February) and Twitter data (28th of February) accounts to one full day for each respective dataset. Next the training and testing of the Naive Bayes classifier is aimed to take up eight days in order that, if the algorithm categorises poorly at first, actions may be executed in order to boost the performance to a sufficient degree. The succeeding task of establishing a sentiment dictionary will occupy three days, which completes "Phase 1". The keyword extraction and the data cleaning of raw data as well as the removal of RAKE candidate keywords will be exercised and assessed for seven days. Determining the number of topics in the following will last for one day. Next the calculation of the aggregated document-topics matrix, aggregated topic-sentiment-keyword matrix and lastly the computation of the opportunity scores will span over three days. The evaluation of the topics' opportunities and identification of the advancement location shortly follows on the next day. Lastly the part of visualising the data occupies seven days in the schedule from the 17th of March to the 24th of March, due to the elevated standards in presenting data for SMEs in order to facilitate the fast extraction of information as well as for documenting project milestones.

Time buffers for experimentation (25th of March to the 1st of April), for reviewing the literature (17th of April to the 22nd of April) and for writing the master dissertation are also implemented, in order to provide extra time for particular tasks, if required. The evaluation of supplemental academic literature is conducted between the 2nd of April and the 16th of April in order to set the master dissertation results into perspective by contrasting it to similar aims and approaches. Finally the writing of the master dissertation is projected to last for 31 days.

Graphic 4:
The Master Dissertation Project Schedule

3.) References

Ahmad, S., Ahmad, N. and Bakar, A. (2018) 'Reflections of entrepreneurs of small and medium-sized enterprises concerning the adoption of social media and its impact on performance outcomes: Evidence from the UAE', *Telematics and Informatics*, 35(1), pp. 6-17. doi: https://doi.org/10.1016/j.tele.2017.09.006.

Barragán-Landy, M. et al. (2020) 'A Proposed Representative Sampling Methodology', *European Conference on Research Methodology for Business and Management Studies*. Aveiro (Portugal), 18th to 19th of June in 2020. Reading (UK): Academic Conferences International Limited, pp. 8-17

Blagus, N. and Zitnik, S. (2018) 'Social media comparison and analysis: The best data source for research?', *12th International Conference on Research Challenges in Information Science (RCIS)*, Nantes (France), 29th to 31st of May in 2018. New Jersey: Institute of Electrical and Electronics Engineers, pp. 1-10.

Burch et al. (2019) 'Finding the outliers in scan path data', *ETRA '19: Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. Denver (USA), 25th to 28th of June in 2019. New York (USA): Association for Computing Machinery, pp. 1-5. doi: https://doi-org.ezproxy.uwl.ac.uk/10.1145/3317958.3318225

Davras, Ö. and Caber, M. (2019) 'Analysis of hotel services by their symmetric and asymmetric effects on overall customer satisfaction: A comparison of market segments', *International Journal of Hospitality Management*, 81, pp. 83-93. doi: https://doi.org/10.1016/j.ijhm.2019.03.003.

Deloitte (2019) *The performance of Small and Medium Sized Businesses in a digital world: A report for the Connected Commerce Council 2019*. Available at: https://www2.deloitte.com/content/dam/Deloitte/es/Documents/Consultoria/The-performance-of-SMBs-in-digital-world.pdf (Accessed: 24th of January in 2021).

Dong, J. and Yang, C. (2020) 'Business value of big data analytics: A systems-theoretic approach and empirical test', Information & Management, 57 (1), pp. 1-9, doi: https://doi-org.ezproxy.uwl.ac.uk/10.1016/j.im.2018.11.001.

Elzayady, H., Badran, K. and Salama, G. (2018) 'Sentiment Analysis on Twitter Data using Apache Spark Framework', *13th International Conference on Computer Engineering and Systems (ICCES)*. Cairo, Egypt, 18th to 19th of December in 2018. New Jersey: Institute of Electrical and Electronics Engineers, pp. 171-176.

Etikan, I. and Bala, K. (2017) 'Sampling and Sampling Methods', *Biometrics & Biostatistics International Journal*, 5 (6), pp. 215-217. doi: 10.15406/bbij.2017.05.00149.

Farizah Ibrahim, N. and Wang, X. (2019) ‘A text analytics approach for online retailing service improvement: Evidence from Twitter’, *Decision Support Systems*, 121, pp. 37-50. doi: https://doi.org/10.1016/j.dss.2019.03.002.

Gallagher, M. and Trendafilov, R. (2018) ‘R vs: Python: Ease of use and numerical accuracy’, *Journal of Business and Accounting*, 11 (1), pp. 117-126.

Ganiger, S. and Rajashekharaiah, K. (2018) ‘Comparative Study on Keyword Extraction Algorithms for Single Extractive Document’, *Second International Conference on Intelligent Computing and Control Systems (ICICCS)*. Madurai (India), 14th to 15th of June in 2018. New Jersey: Institute of Electrical and Electronics Engineers, pp. 1284-1287.

Hu, G. et al. (2017) ‘Analyzing users’ sentiment towards popular consumer industries and brands on Twitter’, *IEEE International Conference on Data Mining Workshops*. New Orleans, LA, USA, 18th to 21st of November in 2017. New Jersey: Institute of Electrical and Electronics Engineers, pp. 381-388.

Jain, K. and Kaushal S. (2018) ‘A comparative study of Machine Learning and Deep Learning Techniques for Sentiment Analysis’, *7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*. Noida (India), 29th to 31st of August in 2018. New Jersey: Institute of Electrical and Electronics Engineers, pp. 483-487.

Jeong, B. and Yoon, J. and Lee, J. (2019) ‘Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis’, *International Journal of Information Management*, 48, pp. 280-290. doi: https://doi.org/10.1016/j.ijinfomgt.2017.09.009.

Kirilenko et al. (2018) ‘Automated Sentiment Analysis in Tourism: Comparison of Approaches’, *Journal of Travel Research*, 57 (8), pp. 1012-1025. doi: https://doi-org.ezproxy.uwl.ac.uk/10.1177%2F0047287517729757.

Koiranen, I. et al. (2019) *Changing patterns of social media use? A population-level study of Finland*. Available at: https://link.springer.com/article/10.1007/s10209-019-00654-1 (Accessed: 29th of October in 2020).

Kritiyanti, D. et al. (2018) ‘Comparison of SVM & Naïve Bayes Algorithm for Sentiment Analysis Toward West Java Governor Candidate Period 2018-2023 Based on Public Opinion on Twitter’, *6th International Conference on Cyber and IT Service Management (CITSM*). Parapat (Indonesia), 7th to 9th of August in 2018. New Jersey: Institute of Electrical and Electronics Engineers, pp. 1-6.

Krouska, A. and Troussas, C. and Virvou, M. (2017) ‘Comparative Evaluation of Algorithms for Sentiment Analysis over Social Networking Services’, *Journal of Universal Computer Science*, 23 (8), pp. 755-768.

Kunal, S. et al. (2018) ‘Textual Dissection Of Live Twitter Reviews Using Naive Bayes’, *Procedia Computer Science*, 132, pp. 307-313. doi: https://doi.org/10.1016/j.procs.2018.05.182.

Lee, I. (2018) ‘Social media analytics for enterprises: Typology, methods, and processes’, *Business Horizons*, 61 (2), pp. 199-210. doi: https://doi-org.ezproxy.uwl.ac.uk/10.1016/j.bushor.2017.11.002.

Li, Y. and Fleyeh H. (2018) ‘Twitter Sentiment Analysis of New IKEA Stores Using Machine Learning’, *International Conference on Computer and Applications (ICCA)*. Beirut, Lebanon, 25th to 26th of August in 2018. New Jersey: Institute of Electrical and Electronics Engineers, pp. 4-11.

Li, Z. and Li, R. and Jin, G. (2020) ‘Sentiment Analysis of Danmaku Videos Based on Naïve Bayes and Sentiment Dictionary’, *IEEE Access*, 8, pp. 75073-75084. doi: https://doi.org/10.1109/ACCESS.2020.2986582.

Liu, X. and Burns, A. and Hou, Y. (2017) ‘An Investigation of Brand-Related User-Generated Content on Twitter’, *Journal of Advertising*, 46 (2), pp. 236-247. doi: 10.1080/00913367.2017.1297273.

Liu, Y. et al. (2020) ‘Cloud-based big data analytics for customer insight-driven design innovation in SMEs’, *International Journal of Information Management*, 51, pp. 1-12, doi: https://doi-org.ezproxy.uwl.ac.uk/10.1016/j.ijinfomgt.2019.11.002.

Maroufkhani, P. et al. (2020) ‘Big data analytics adoption: determinants and performances among small to medium-sized enterprises’, *International Journal of Information Management*, 54, pp.1-15. doi: https://doi-org.ezproxy.uwl.ac.uk/10.1016/j.ijinfomgt.2020.102190.

Molinillo, S. et al. (2019) ’Smart city communication via social media: Analysing residents’ and visitors’ engagement’, *Cities*, 94, pp. 247-255. doi: https://doi.org/10.1016/j.cities.2019.06.003.

Mukhtar, N. and Abid Khan, M. and Chiragh, N. (2018) ‘Lexicon-based approach outperforms Supervised Machine Learning approach for Urdu Sentiment Analysis in multiple domains’, *Telematics and Informatics*, 35 (8), pp. 2173-2183. doi: https://doi-org.ezproxy.uwl.ac.uk/10.1016/j.tele.2018.08.003.

Nakayama, M. and Wan, Y. (2019) ’The cultural impact on social commerce: A sentiment analysis on Yelp ethnic restaurant reviews’, *Information & Management*, 56(2), pp. 271-279. doi: https://doi.org/10.1016/j.im.2018.09.004.

Ofcom (2017) *Adults' media use and attitudes Report 2017*. Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0020/102755/adults-media-use-attitudes-2017.pdf (Accessed: 29th of October in 2020).

Office of Management and Budget (2017) *North American Industry Classification System.* Available at: https://www.census.gov/eos/www/naics/2017NAICS/2017_NAICS_Manual.pdf (Accessed: 26th of January in 2021)

Rajabion, L. (2018) 'Application and Adoption of Big Data Technologies in SMEs', *International Conference on Computational Science and Computational Intelligence (CSCI)*. Las Vegas (USA), 12th to 14th of December in 2018. New Jersey: Institute of Electrical and Electronics Engineers, pp. 1133-1135.

Resyanto, F. and Sibaroni, Y. and Romadhony, A. (2019), 'Choosing The Most Optimum Text Preprocessing Method for Sentiment Analysis: Case: iPhone Tweets', *Fourth International Conference on Informatics and Computing (ICIC)*, Semarang (Indonesia), 16th to 17th of October in 2019. New Jersey: Institute of Electrical and Electronics Engineers, pp. 1-5.

Rustam, F. et al. (2019) 'Tweets Classification on the Base of Sentiments for US Airline Companies', *Entropy*, 21 (11), pp. 1078-1099. doi: https://doi.org/10.3390/e21111078.

Singh, R. and Verma, H. (2020) 'Effective Parallel Processing Social Media Analytics Framework', *Journal of King Saud University - Computer and Information Sciences*, Not yet assigned to volumes/issues, pp. 1-11. doi: https://doi.org/10.1016/j.jksuci.2020.04.019.

Statista Research Department (2016) *Digital presence of U.S. SMEs 2016*. Available at: https://www.statista.com/statistics/680426/sme-owners-digitial-presence-usa/ (Accessed: 27th of January in 2021)

Twitter (2021a) *Twitter's search endpoints*. Available at: https://developer.twitter.com/en/docs/twitter-api/search-overview (Accessed: 16th of January in 2021).

Twitter (2021b) *Premium search APIs*. Available at: https://developer.twitter.com/en/docs/twitter-api/premium/search-api/api-reference/premium-search#Pagination (Accessed: 29th of January in 2021).

Ulwick, A. (2005) '*What Customers Want: Using Outcome-Driven Innovation to Create Breakthrough Products and Services'*. New York City (USA): McGraw-Hill Education Ltd.

Vayansky, I. and Kumar, S. (2020) ‘A review of topic modeling methods’, *Informations Systems*, 94, pp. 1-15. doi: https://doi-org.ezproxy.uwl.ac.uk/10.1016/j.is.2020.101582.

Wang, B. et al. (2014) ‘Identifying technological topics and institution-topic distribution probability for patent competitive intelligence analysis: a case study in LTE technology’, *Scientometrics*, 101, pp. 685-704.

Yelp (2021) */business/{id}/reviews*. Available at: https://www.yelp.com/developers/documentation/v3/business_reviews (Accessed: 16th of January in 2021).