Big Data opinion mining on Social Media platforms: A review on recent academic literature

Table of Contents

0.) Abstract

Social media platforms exhibit rich data in terms of perspectives towards products, services and brands. Through the literature review three key topics were encountered, which involve first the interpretation of the social media users' behaviour by corporate or state owned entities and the trade-offs these organisations face, second the benefits of opinion mining for assisting decision makers to invent or improve products, services and brands suited for a target group and third the scaling and optimal utilisation requirements of the IT-architecture for being capable to examine even tremendous volumes of social media data within a short processing time period and within low operational cost limits. Furthermore gaps have been detected first in the area of applying the opportunity algorithm (a social media analysis technique) on social media feedback addressed towards the service industry, second in the academic awareness for social media exploitation opportunities for small to medium sized enterprises (SMEs), third in the research examining social media data outputted by micro-blogging platforms for the retail sector, fourthly in the successful application of social media strategies in the context of smart cities directing online attention to local brands and the prestige of the smart city. Last but not least research questions were formulated in order to analyse how SMEs may bridge the asymmetry between their scarce resources and the tremendous quantity of social media data in order to derive beneficial knowledge for advancing their multi-functional service.

1.) Introduction

Social media (SM) presents a repository of tremendous extend to harvest data for companies, governments and state entities as well as for researchers in order to attain more refined degrees of accuracy in evaluating public opinion. To illustrate, SM networks managed in the recent past to attract significant portions of people throughout society to join ((Koiranen et al., 2019), (Ofcom, 2017)) and participate in form of ratings, reviews and recommendations at their respective platforms (Ahmad and Ahmad and Bakar, 2018). Therefore it should come as no surprise that companies followed potential and existing clients to these networks for purposes like promoting their businesses and their particular brands (Ahmad and Ahmad and Bakar, 2018) accompanied by low budget expenses for marketing asserted by Brajos-Gomes and Benitez-Amado and Llorens-Montes (2015, cited in Ahmad and Ahmad and Bakar, 2018) or for gaining an enhanced understanding of customers' perspectives in order to estimate with an increased certainty the trends itself and their relevance ((Ahmad and Ahmad and Bakar, 2018), (Li and Fleyeh, 2018)). Likewise governments and politicians may benefit from conducting a SM data driven approach in order to derive support for decision making (Singh and Verma, 2020) for instance when planning to adjust incentives for companies to settle according to the population's opinion (Li and Fleyeh, 2018) resulting in being recognised as a capable leader.

The scope of this literature review comprises three relevant and generic themes for decision makers to take into account when dealing with SM data. First the interpretation of SM users' perception on a particular topic considering their social, cultural and demographic influences ((Koiranen et al., 2019), (Ofcom, 2017), (Nakayama and Wan, 2019), (Molinillo et al., 2019), (Li and Fleyeh, 2018), (Ahmad and Ahmad and Bakar, 2018)), second the beneficial knowledge extracted from opinion mining in terms of product, service and brand development ((Jeong and Yoon and Lee, 2019), (Farizah Ibrahim and Wang, 2019), (Hu et al., 2017)) and third the scalability in terms of computing power and memory usage in order to cope with the vast amount of data ((Singh and Verma, 2020), (Elzayady and Badran and Salama, 2018)) in order to categorise sentiment in a timely fashion (Kunal et al., 2018).

Despite the essentiality of SM data analysis for decision makers outlined above, vital limitations still remain uncovered. Concerns range from the lack of research for the impact of SM on small to medium sized enterprises (SMEs) in developed countries (Ahmad and Ahmad and Bakar, 2018) over the scarcity of academic awareness towards the exploitation options of data from microblogging platforms like Twitter for the field of the retail industry (Farizah Ibrahim and Wang, 2019) to the undeveloped capabilities of smart cities to spark online engagement regarding its local brands and its reputation for visitors as for inhabitants (Molinillo et al., 2019).

For the purpose of providing a more detailed overview over the three generic themes this paper will subsume, analyse and connect first for each individual theme and then perform this process again with the difference of performing it in an overarching manner across the specified themes for the section of conclusions, which precedes the final part of this literature review i.e. references.

2.) Crucial aspects of Social Media and how to consider to them

In Table 1, the individual research articles will be listed together with their main aims, key findings, advantages and drawbacks as well as their assigned category according to one of the themes mentioned above. The abbreviation NLP is used for the term Natural Language Processing.

Researchers	Aim(s)	SM Theme	Key Findings	Ad-vantages	Drawbacks
Li and Fleyeh	• Collecting, evaluating and estimating local sentiment, awareness as well as related topics regarding novel IKEA shop openings by utilising NLP techniques	• Percep-tion	• The General Linear Model with elastic net penalty was outperforming the remaining techniques • Emoticons were seen as beneficial for automatic sentiment categorisation of each Swedish tweet, whereas English tweets were classified via the lexicon method • Larger cities linked topics like traffic and environment to IKEA • Small-scale cities focused more on lifestyle and room renovation in conjunction to the keyword IKEA	• F1 and AUC scores were measur-ed to control for a balanc-ed senti-ment distribu-tion in the dataset • Comput-ed Pearson cor-relation coeffici-ents between the key-word and each re-maining word to infer the associa-tion to the keyword	• The tweets' probability of an evenly distributed sentiment dataset was slightly biased towards the positive segment, which ranged between 0.6-1 in the re-searchers' approach as compared to their neutral (0.3-0.6) and negative (0-0.3) sentiment ranges

Researchers	Aim(s)	SM Theme	Key Findings	Ad-vantages	Drawbacks
Elzayady and Badran and Salama	• Proposing a scalable architecture for an efficient sentiment prediction technique	• Computing Speed and time relevance of information	• Naive Bayes and Logistic Regression were achieving similar results, while decision trees was lacking in performance	• Apache Spark was characterised as aptly in particular for Machine learning algorithms by the academics	• 70% of training data used while conducting 5-fold cross-validation
Kunal et al.	• Create an accurate tool for identifying at an early stage extremist textual content on live Twitter data	• Computing Speed and time relevance of information	• The technique of Naive Bayes attained a significantly higher degree of precision than the method of Decision Trees • This approach in general generates live sentiment corresponding to the live data and is topic-independent as well as language-independent	• Importing the Regular Expression library for data cleaning • Naive Bayes Model attains an F1 score of about 0.9 while decision trees are achieving 0.65 for the same category	• Textblob needs to be adjusted by an extension for each different language, respectively

Researchers	Aim(s)	SM Theme	Key Findings	Ad-vantages	Drawbacks
Koiranen et al.	• Conclude trends in the develop-ment of social media usage in Finland from the year 2008 to 2016 across a range of socio-demographic groups	• Percep-tion	• Significant diminished social effect of SM • It is more probable in 2016 to have a SM account than in the year 2008 independent of the socio-demographic background • An asymmetric technology diffusion in terms of SM could be identified between distinct age groups • Education was not a significant factor for being attracted to brands • Elevated brand connectedness was only seen for the age brackets between 16 and 44	• Data is repre-sentative for Finland • The trends may be used as estima-tion for future usage behav-iour in less digital-ised countries	• The participants were non-identical between the monitored time span • The data and therefore the trends are likely not exemplary for entire Europe, which has been stated by the academics

Researchers	Aim(s)	SM Theme	Key Findings	Ad-vantages	Drawbacks
Molinillo et al.	• Evaluating online engagement features of SM for visitors and inhabitants of ten Spanish smart cities	• Percep-tion	• People interacted the most through likes, secondly through shares and thirdly via comments • Passively one-sided communication (likes, shares and retweets) is the most prevalent form of SM engagement across Spain (also claimed by the researchers for the EU) • Instagram had the foremost user engagement • Facebook was attracting more likes and comments than Twitter • Through Twitter more shares (retweets) have been conducted than via Facebook	• For cities there is a trade-off between user engage-ment and an extensive number of followers of the SM channel • Identified that the smart cities were just making basic use of the SM potential for their own local brands and for their image	• Despite the research was aiming for visitors as well as inhabitants, only SM channels of visitors were selected for data processing • The academics utilised commodity spread-sheet software, which is unsuitable for scalable data analysis • Data collection from a range of 10 Spanish smart cities

Researchers	Aim(s)	SM Theme	Key Findings	Ad-vantages	Drawbacks
Jeong and Yoon and Lee	• Identifying latent product features with the highest unserved demand potential	• Product, service and brand develop-ment	• Future directions for demand can be inferred • The opportunity algorithm acts like a compass for multi-functional products so that producers may include or enhance the most desired feature(s) by customers of a current product • This method is able for live monitoring • Capable to assist in the design stage of novel products	• RAKE keyword extrac-tion model is inte-grated	• Limited to a multi-functional product • Not applied to multiple products • Not exer-cised on services or product-service systems • Reddit posts are not the same as customer reviews • About 50% of the dataset was cleaned • Discarded circa 90% of Rake keywords • Topics were manually assigned

Researchers	Aim(s)	SM Theme	Key Findings	Ad-vantages	Drawbacks
Singh and Verma	• Operating a scalable and economically affordable approach for measuring big data techniques in terms of sentiment prediction accuracy and processing time on a tremendous amount of real-time data	• Comput-ing Speed and time rele-vance of informa-tion	• Input data without textual content was not significant • Light gradient boosting decision tree model attained the highest F1 score of all approaches • Rank of importance for tweaking estimation results is: First to include URL data, second to integrate Image data and third to incorporate Tweet's author's metadata • Faster processing time and more efficient CPU and memory usage could be attained via a multi-threading architecture, the more threads were added	• Utilising Twitter image and URL inform-ation	• The proposed architec-ture was not compared to a purely distributed or a distributed and parallel approach in terms of computa-tional speed and memory usage • AYLIEN API for evaluating images and URLs was limiting speed improve-ments especially when multi-threading was exercised

Researchers	Aim(s)	SM Theme	Key Findings	Ad-vantages	Drawbacks
Ahmad and Ahmad and Bakar	• Getting insights from SMEs about their SM usage and its benefit in the UAE	• Percep-tion	• SMEs have limited resources • SMEs goals were to attract either new client segments in existing markets or to gather novel customers from new regions across all companies • Another aim by SMEs was to enhance the customer relationship	• A need for SM analysis by SMEs was detected to monitor current customer demand, design new products and services, plus for grounded decision support	• The number of companies asked was 7 all from one single country
Hu et al.	• Discovering the variations of: Sentiments in different industries, between sentiment in brands of the same industry as well as of sentiment towards topics within distinct industries	• Product, service and brand develop-ment	• Positive sentiment for the producing and negative sentiment was identified for the service sector • Topic sentiments are conditional to the industry • Sentiments towards brands were amplified compared to the general state	• Brands with a tremen-dous amount of online-engage-ment exhibit more neutral senti-ment	Manually assigning each brand to 1 or more industries, which is an area exposed to human error

Researchers	Aim(s)	SM Theme	Key Findings	Ad-vantages	Drawbacks
Ofcom	• Display SM usage behaviour for socio-demographic groups in the UK between at least the time span of 2012 to 2016	• Percep-tion	• More novel participants were on SM platforms across age groups • Distinct social grade groups have a distinct preference distribution for SM platforms • The youngest age group depicted the most diverse preference distribution for key SM platforms	• Interest-ing activities (clicking on adds, etc.) could be quantifi-ed in terms of percenta ges of parti-cipants perform-ing at least once the action	• In part multiple activities (e.g. posting comments or sharing photos and videos) were classified as one distinct under-taking (e.g. clicking ads)
Farizah Ibrahim and Wang	• Exploit Twitter user data for discovering crucial topics of negative sentiment among online retailers	• Product, service and brand develop-ment	• Delivery and customer service contained to an enormous degree negative sentiment • Network analysis depicted the link between LDA topics (i.e. delivery timeliness and product availability)	• 2 new topics (Online-engage-ment, In-Store experi-ence) were discover-ed through LDA topic modelling	• Assigning topics manually • Some pre-known topics from the literature were not identified • 17% of tweets contained a non-neutral sentiment

Researchers	Aim(s)	SM Theme	Key Findings	Ad-vantages	Drawbacks
Nakayama and Wan	• Examining to what degree culture has an impact on social commerce on the example of Japanese restaurants	• Percep-tion	• Ethnic culture effects the evaluation of customer reviews and review help-fulness • Food quality was perceived as the major important restaurant characteristic independent of culture • The three remaining characteristics were depending on the culture • Reviews including food quality were rated as the most helpful across sentiment and across cultures • The residual properties were varying across sentiments and across cultures	• Use of associa-tion for identifi-cation of senti-ment words/ phrases with searched key terms (Japan-ese starters) via IBM Watson Content Analytics (WCA) 11.0 • WCA is suited e-specially for a language	• WCA is limited in the amount of available languages • The sentiment terms have to be manually assigned to a topic category, which implies subjectivity to some extend • More than 37% of sentiment phrases were allotted to the un-known sentiment class

Table 1:
Overview and categorisation of the reviewed academic literature

2.1.) Perceptions depending on social, cultural and demographic backgrounds

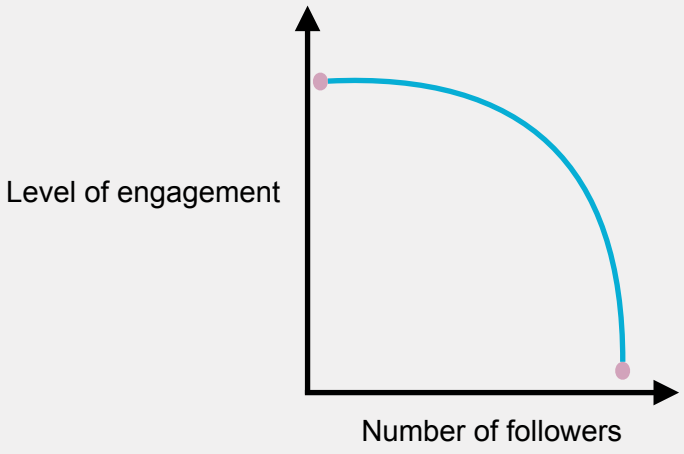
Each decision making entity whether it is owned by the government or a private corporation has to clarify its SM goal(s) for which an appropriate SM strategy is required in order to successfully achieve these aims. As described in Table 1 under the SM rubric perception, the assigned academics' contributions were either enhancing the understanding of potential target groups ((Li and Fleyeh, 2018), (Nakayama and Wan, 2019), (Koiranen et al., 2019), (Ofcom, 2017)) or bringing light to the area of trade-offs from the standpoint of a decision maker ((Ahmad and Ahmad and Bakar, 2018), (Molinillo et al., 2019)).

For the former category, the scholars ((Li and Fleyeh, 2018), (Nakayama and Wan, 2019)) could draw conclusions from the locally and culturally collected dataset, for which sentiment towards either Japanese restaurants or the room inventory store IKEA was evaluated in order to infer the local populations' degree of satisfaction. In addition Li and Fleyeh (2018) were also measuring the magnitude of people's awareness over a specific time period before, at and after the new IKEA store was commencing its business operations, which may be beneficial for predicting future trends of novel store openings of this company. Interestingly both author groups were applying the same method of measuring correlation and thereby deducing association ((Li and Fleyeh, 2018), (Nakayama and Wan, 2019)) for distinct purposes. According to Table 1 the goal of Li and Fleyeh (2018) was to identify related topics for the keyword IKEA and to compare these associations between cities, while Nakayama and Wan (2019) were aiming to extract relevant sentiment keywords and key phrases linked to Japanese starters. Albeit Nakayama and Wan (2019) could not classify at least 37% of the sentiment phrases and categorised the remainder of the data by potentially subjective human beings, they managed to draw the picture that food quality was the prior concern across cultures in the positive as well as in the negative direction. Furthermore Nakayama and Wan (2019) ranked and outlined the cultural divergence in terms of negative and positive sentiment for the outstanding features in order to conclude, which features of restaurant reviews and therewith restaurants are perceived as crucial for Japanese and Western customers respectively. In contrast Li and Fleyeh (2018) were not labelling sentiments like Nakayama and Wan (2019) with the help of the rating score, but partially by the lexicon approach for English or for the language of Swedish, of which no well developed sentiment lexicon existed, they utilised emoticons to automatically assign slightly biased emotional value to textual content as can be seen in more detail in Table 1.

The remaining intellectuals ((Koiranen et al., 2019), (Ofcom, 2017)), who were scrutinising the perspectives of socio-demographically distinct user groups through representative surveys, demonstrated similarly, that for their examined countries, user numbers of SM platforms have risen in the time period of the years 2012 to 2016 independent of age groups for the UK (Ofcom, 2017) and between the years 2008 and 2016 independent of all socio-demographic characteristics for Finland (Koiranen et al., 2019). Additionally Koiranen et al. (2019) indicated that participants under the age of 45 were significantly more eager to follow brands, while the youngest bracket in the study by Ofcom (2017) was showing a more diverse preference distribution when being asked for its main platform for SM activities in comparison to the remaining age groups. However both academic articles were studying different activities among the observed population. While Ofcom (2017)

conducted research concerning the preferred SM network among individual social grade groups and recorded activities like clicking on adds, communicating thoughts to strangers and looking at post without interacting to them, Koiranen et al. (2019) was measuring the sunken social effect of SM over the years, the increase in asymmetric SM diffusion across the distinct age groups, which in particular meant that people of younger age were more probable to have an account on a SM platform compared to older generations. In contrast the limitations of the survey by Koiranen et al. (2019) were that, the surveyed partakers were likely to be not identical between the observation periods, while the survey maintained the characteristic of being representative. Furthermore Finland is not being an excellent epitome of Europe's activities although the country is a forerunner for less technological developed countries in terms of SM usage trends according to Koiranen et al. (2019). Alternatively the drawback of the study by Ofcom, (2017) according to Table 1 consisted in part out of distinct actions, which were asked and collected in a conjunctive manner for e.g. in the category of posting comments or sharing photos and videos, which diluted the explanatory power of these findings.

Last but not least the reasonings drawn from the scholars, who inspected the trade-offs of decision makers, were that there is first a trade-off between invested (monetary, temporal and human capital) resources and the resulting (brand and company) awareness in the case of SMEs (Ahmad and Ahmad and Bakar, 2018) and there is second in the context of smart cities the trade-off between the level of online engagement (likes, shares and comments) and the number of followers (Molinillo et al., 2019). The latter trade-off is illustrated in Graphic 1, in which a city can maximise its influence by taking one of two directions. Either the city chooses a maximum level of online engagement by neglecting the number of followers or it aims for a tremendous amount of followers by paying with a low degree of online engagement as it is concluded in the academic essay by Molinillo et al. (2019) and here portrayed with the two purple dots. Therefore the blue trade-off curve assumes in Graphic 1 that it is more favourable to aim for just one of the solutions (i.e. one of the purple dots) than to strategically target a mix of the two dimensions, which is consistent with the writings by Molinillo et al. (2019). Combining this knowledge with the qualitative findings by Ahmad and Ahmad and Bakar (2018) described in Table 1, one may notice that the SMEs should therefore consider whether it is more relevant and suitable to them to increase the level of engagement for example by enhancing the customer relationship or to pursue a growth strategy in terms of attracting higher follower numbers. Despite this conclusion, one should take into account that quantitative studies may expose problems when they are generalised, although they may provide a



Graphic 1:
Trade-off between
online engagement and the number of followers

good indication for further research. On the other hand Ahmad and Ahmad and Bakar (2018) recognised in their essay that there is a call for SM data analysis by SMEs for the purposes of first monitoring existing customer demand, second for the development of novel products and services and third for a more stable basis for decision making. The same has to be considered of the article written by Molinillo et al. (2019). While these intellectuals discovered the online engagement for the different city's visitor SM channels like it is outlined in Table 1, their data analysis was just confined on ten smart cities located in Spain. In addition Molinillo et al. (2019) were failing in their aim to evaluate visitor as well as inhabitants' online engagement since they were only examining the SM channels addressed to visitors and therefore could not differentiate engagement by visitors or by inhabitants. Last but not least the named academics operated their data analysis with a spreadsheet application, which is unsuitable in terms of scaling architecture when being challenged with the big data nature of smart cities in future experiments. However, both research articles revealed relevant gaps for instance in SM studies for SMEs in developed countries (Ahmad and Ahmad and Bakar, 2018) as well as in the opportunity to improve online engagement in the realm of smart cities (Molinillo et al., 2019).

Given the above, one is able to identify the advantages of locally confined SM big data analysis, the benefits of conclusions derived from SM usage studies and the enhanced strategy decision support obtained from paying attention to decision makers' trade-offs especially in the context of SMEs, under the caveat that the firm's respective aims for online engagement or for an extensive amount of SM audience and their locally targeted customer group(s) are properly considered.

2.2.) Insights about products, services and brands through Big Data analysis

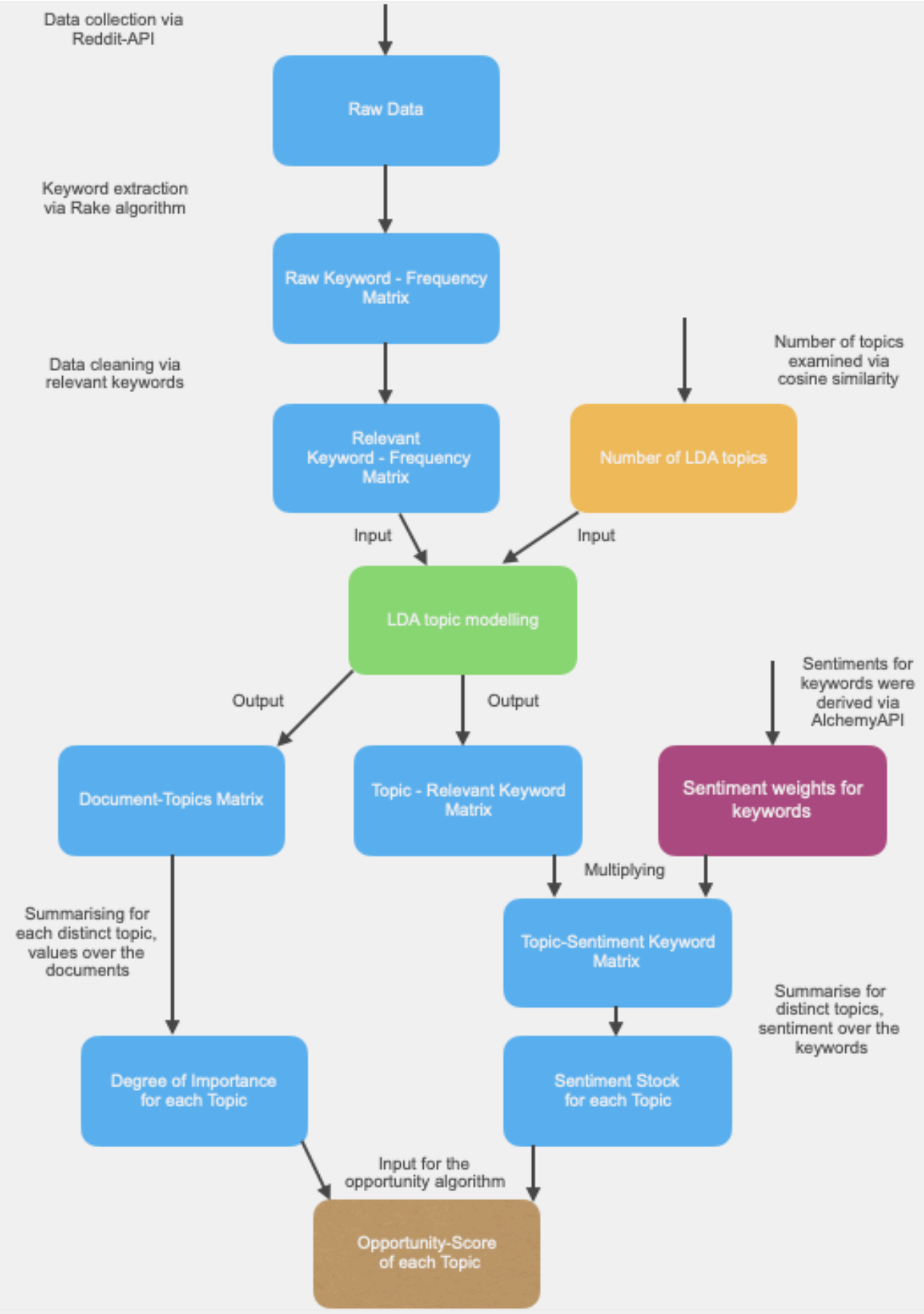
After comprehending how the SM user thinks in regard to companies, for what activities she or he is utilising SM platforms and which trade-offs decision makers face, it is crucial to understand in which service or product directions branch-specific firms have the potential to refine their current or future portfolio towards the clients' desires. Furthermore before applying SM big data analytics in order to mine customer or public opinions, it may be beneficial to prioritise the identification process due to time or computing constraints. For this reason the findings by Hu et al. (2017) can be a suitable support, which were concluding that positive sentiment was mostly directed at producing branches of the economy, whereas negative sentiment was mainly addressed towards the service sector industry, for which Farizah Ibrahim and Wang (2019) discovered for the subsection of online retail, that delivery as well as customer service were exposed to a tremendous level of negativity too while the product characteristic of product price was receiving moderate negative sentiment. Therefore the opportunity algorithm described and applied on just a single multi-functional product by Jeong and Yoon and Lee (2019) (designed to support the development of current or novel product features as indicated in Table 1) seems to be very beneficial for serving the significant gap in analysing multi-functional services since a low satisfaction score for a topic, which is defined as an enormous amount of negative sentiment towards a topic, is occurring in the service sector more frequently than for product features as indicated by Hu et al. (2017) as well as by Farizah Ibrahim and Wang (2019). The process of how the opportunity score is calculated by Jeong and Yoon and Lee (2019) may be withdrawn from Graphic 2, while the definition of the opportunity score is provided in Formula 1:

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

Formula 1:

The opportunity formula according to Jeong and Yoon and Lee (2019)

In more detail, LDA (Latent Dirichlet Allocation) topic modelling, which is one of the core components of the opportunity algorithm process, is also conducted in the approach by Farizah Ibrahim and Wang (2019) leading to the discovery of two novel topics, which were not found by the academics in the literature. In sharp contrast Farizah Ibrahim and Wang (2019) admitted the drawback of not discovering some well-known topics, which are outlined by their read literature as is written in Table 1. A further critique concerning the LDA method which was performed by the intellectuals Jeong and Yoon and Lee (2019) as well as by Farizah Ibrahim and Wang (2019) is the exposure to subjective human error in the step of assigning topic labels to groups of (key)words. However one must also determine a suitable number of topics for the LDA algorithm as input, which Jeong and Yoon and Lee (2019) solved via measuring the least amount of cosine similarity between topics, whereas alternatively Farizah Ibrahim and Wang (2019) performed Gibbs sampling. In addition Farizah Ibrahim and Wang (2019) was performing a network analysis



Graphic 2:

The opportunity algorithm process according to Jeong and Yoon and Lee (2019)

between their extracted LDA topics, upon which the scholars inferred the significant relationship between delivery timeliness and product availability.

The two remaining core components of the opportunity algorithm concern the sentiment integration and the keyword extraction as it is depicted in Graphic 2 in detail for the approach by Jeong and Yoon and Lee (2019). Therefore the use of sentiment and keyword extraction tools is evaluated among Jeong and Yoon and Lee (2019), Farizah Ibrahim and Wang (2019) as well as Nakayama and Wan (2019), excluding the keyword extraction characteristic of the LDA topic modelling algorithm Mallet by Farizah Ibrahim and Wang (2019) due to the lack of provided information concerning the ratio of omitted to relevant data. Moreover Jeong and Yoon and Lee (2019) applied RAKE (Rapid keyword extraction algorithm) and requested for sentiment classification the Alchemy API like depicted in Graphic 2, while Nakayama and Wan (2019) operated with IBM WCA for keyword or key phrase extraction and then classified sentiment with the help of the Yelp rating, so that reviews with 5-4 stars were considered positive and reviews from 1-2 stars were assessed as negative, whereas last but not least Farizah Ibrahim and Wang (2019) categorised sentiment via SentiStrength. The techniques integrated in SentiStrength were able to assign 17% of the data to a non-neutral sentiment class, while Nakayama and Wan (2019) were able to group most of the reviews correctly except for the 3 star rating ones, which represented neutral rating. In addition IBM WCA could just assign at the maximum 69.4% to the latter researchers' topics, which were connected to keywords or key phrases correlated to Japanese starters, while Jeong and Yoon and Lee (2019) implemented Alchemy API by IBM, which they claim to be very accurate, but through conducting tremendous data cleaning, the raw keywords obtained through the RAKE method, shrunk to just about 10 % of essential keywords and thereby the corpus was more than halved, because documents without essential keywords were omitted.

Further findings relating to sentiment by Hu et al. (2017) indicated, that sentiment expressed for brands fluctuated more than the generic sentiment for positive as well as for negative amplitudes. Furthermore Hu et al. (2017) found out in their research that there is more neutral sentiment for brands with a large-scale followership as presented in Table 1, which is consistent with the trade-off brought into the discussion by Molinillo et al. (2019) in the last section.

Ultimately it can be inferred that entities examining SM data are advised to scrutinise with a higher priority positive sentiment for the manufacturing branch and negative sentiment for the service sector under the support of big data techniques in order to adapt their supply of products and/or services to current trends of customers' demand resulting in an increased creation of wealth for both sides.

2.3.) The computing speed and time relevance of information

At the stage of the implementation of a particular exploitative algorithm for SM data, there is the question of how to design the underlying architecture in terms of affordability and in terms of processing speed while handling the enormous amount of input data which may be analysed in a short time window. Therefore Singh and Verma (2020) were proposing a parallel computing scheme in which the researchers were utilising threading and a message broker, whereas Elzayady and Badran and Salama (2018) were operating on the parallel and distributed Apache Spark framework. In Graphic 3 the significant outcomes of the scholars respective approaches were set into comparison in order to recognise the advantages that the more a process is threaded Singh and Verma, (2020) or set in parallel Elzayady and Badran and Salama (2018) the lower was the execution time for predictive analysis tasks on SM data and across a range of machine learning (ML) algorithms. In addition Singh and Verma (2020) reported that in their model the CPU as well as the memory utilisation are more efficient the more threads they were operating.

Dataset in amount of tweets	Runtime in s for 1 node	Runtime in s for 2 nodes	Runtime in s for 3 nodes
100.000	ca. 250	ca. 180	ca. 120
In Comparison	1	0.72	0.48
200.000	ca. 570	ca. 380	ca. 290
In Comparison	1	0.67	0.51

Activity	Single threaded design computation time in s	2 threaded design computation time in s	4 threaded design computation time in s
Complete Processing	993.31	573.41	484.06
In Comparison	1	0.58	0.49

Graphic 3:
Main observations by Elzayady and Badran and Salama (2018) in upper table
as well as by Singh and Verma, 2020 in the lower table

Kunal et al. (2018) as well as Elzayady and Badran and Salama (2018) were furthermore discovering that Naive Bayes was outperforming the method of decision trees while Elzayady and Badran and Salama (2018) also stated that the result of Naive Bayes was similar to the one of Logistic Regression. In sharp contrast Singh and Verma (2020) examined three supervised ML techniques namely Support Vector Machines, Long Term Short Memory and Light Gradient Boosting Machine (LGBM), of which the latter one was first derived from decision trees and second

it was achieving the highest F1-score according to Table 1 although the remaining approaches were in very close proximity to the accuracy, precision, recall and F1-scores of LGBM.
More essential conclusions were withdrawn from Singh and Verma (2020), which inferred from their own experiments for the task of predicting sentiments that the texts of tweets are holding an extensive amount of relevance so that in comparison any other combination without this component i.e. text information was categorised as not significant. Furthermore the URL, image data and authors' metadata could be ranked according to their assisting importance in supporting the estimation like it is written down in Table 1 under the subsection key findings by Singh and Verma (2020), while the distinct prediction technique Naive Bayes utilised by Kunal et al. (2018) has been classified as topic and language independent although the text-mining library which is integrated in this particular approach will return more accurate results once the list of stop words is updated to the examined language's specific stop words. In sharp contrast to the advantageous reasoning by Singh and Verma (2020), falls the critique addressed to them of the lack in comparing their architecture to a simple distributed as well as to a distributed and parallel cluster computing design in terms of computing speed plus memory and CPU utilisation. Moreover another drawback in the academic article authored by Elzayady and Badran and Salama (2018) concerns, despite applying 5-fold-cross-validation, the extensive share of the training set as indicated in Table 1, while they were executing in advance the keyword extraction method TF-IDF (Term frequency - Inverse Document frequency), which is one of the competitors of the RAKE technique conducted in Jeong and Yoon and Lee (2019) and portrayed in Graphic 2. Last but not least the similarities between Kunal et al. (2018) and Jeong and Yoon and Lee (2019) became clearer due to Kunal et al. (2018) training Naives Bayes on an external dataset, which due to its size and its word variations is considered to be representative and the latter academics, who were requesting a well trained, but external sentiment classifier for individual keywords as it is depicted in Graphic 2.
It can be concluded that in order to maintain a low processing time when analysing an immense quantity of SM data, a scalable parallel configuration of the underlying design infrastructure is suitable and required also by looking at the condition of staying within a strictly confined budget. Furthermore it can be observed from Table 1 that the examinations of distinct evaluation techniques where partly contradicting each other, while all of the methods were belonging into the group of supervised ML algorithms. Thirdly it has been determined that textual content of tweets was the core component for extracting valuable sentiment information (Singh and Verma, 2020). Last but not least the alternative to the RAKE method, namely the TF-IDF algorithm was introduced to select relevant keywords and the ease of use for operating with an external sentiment classifier, which was developed by a reputable computing company, was highlighted by Jeong and Yoon and Lee (2019), which may support especially SMEs due to their extremely limited resources in analysing at an accelerated pace vast portions of SM data.

3.) Conclusions

By taken all the information into account four clear gaps could be identified. First the absence of the opportunity algorithm's application on the negatively connoted service sector was discovered, second the insufficient research for SM effecting SMEs within developed countries (Ahmad and Ahmad and Bakar, 2018) was highlighted, third the tremendous potential in the retail industry for examining SM micro-blogging data (Farizah Ibrahim and Wang, 2019) was underlined and last but not least the deficiency of proper SM fostering of regional brands and the appropriate depiction of a city's profile in the context of smart cities (Molinillo et al., 2019) was indicated. Furthermore the decision support desire by a SME requires a SM exploitative approach (like the opportunity algorithm), which analyses SM data for the SMEs in order to assist in designing novel services and products, or in order to support the exploration of the prevailing customer demand (Ahmad and Ahmad and Bakar, 2018) and it was therefore categorised as highly relevant due to the giant amount of potentially beneficial SM data (Singh and Verma, 2020) SMEs are also confronted with, while recognising their limited resources (Ahmad and Ahmad and Bakar, 2018).

Three main themes derived from this literature review for decision makers to consider when formulating a SM strategy could be outlined. The first one being the individual perceptions of SM users depending on their socio-demographic and trade-off-conditional situation, the second one comprising the potential for supportive knowledge derivation obtained from SM data for brands (Hu et al., 2017), services (Farizah Ibrahim and Wang, 2019) and products (Jeong and Yoon and Lee, 2019) in order to evolve the features to the clients needs and last but not least the operational and IT-architectural aspect of utilising commodity hardware in the most effective and efficient manner like it is performed under the parallel as well as threaded designs in order to cope with the extensive existing SM user generated data, ((Singh and Verma, 2020), (Elzayady and Badran and Salama, 2018)) and its corresponding time-sensitive sentiment (Kunal et al., 2018). Therefore the following four research questions (RQ) were derived and classified as of great interest for future research:

- RQ1: Is the opportunity algorithm also deployable for detecting important and unsatisfied characteristics and for indicating feature improvement directions for a multi-functional service?

- RQ2: If RQ1 is true, how do opportunity score and the area of optimisation for the features of multi-functional service differ from each other and which characteristics are similar among distinct Western cities?

- RQ3: Does Network analysis support the understanding of the relationships between various returned LDA topics for a city and if this is true in which way?

- RQ4: How much time may be saved through operating parts of the program on a cluster computing architecture compared to a single computer and is the computing time difference between these two designs significant?

Finally the significance of the insights that assigning sentiments to texts in a quick, but precise fashion like it is demonstrated by Li and Fleyeh (2018) with the help of emoticons or by Jeong and Yoon and Lee (2019) under the assistance of external sentiment classifiers as well as the conclusion that the textual content of SM data is the key pillar for extracting sentiment information (Singh and Verma, 2020) should also not be forgotten.

4.) 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.

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.

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.

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.

Jeong, B., 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.

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).

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.

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 2018. New Jersey: Institute of Electrical and Electronics Engineers, pp. 4-11.

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.

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)

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.