Cloud-based Big Data Analytics for Customer Insight-driven Design Innovation in SMEs

Authors: Ying Liu[a],*, Anthony Soroka[b], Liangxiu Han[c], Jin Jian[d], Min Tang[e]

a Institute of Mechanical and Manufacturing Engineering, School of Engineering, Cardiff University, Cardiff CF24 3AA, UK
b Cardiff Business School, Cardiff University, Cardiff CF10 3EU, UK
c School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University,Chester Street, Manchester M1 5GD, UK
d Department of Information Management, School of Government, Beijing Normal University,Beijing 100875, China
e School of Management, Chongqing Technology and Business University, Chongqing 400067,China

Abstract Fast development of IT and ICT facilitate customers to post a large volume of their concerns and expectation online, which are widely accepted to be a valuable resource for product designers. However, it is found that only a small number of small and medium-sized enterprises (SMEs) have capabilities to leverage customer online insights for design innovation, which often demonstrate a significant share in national economies growth. To discover the beneath reasons regarding the barrier that prevent them to make effective utilization, in this study, as a concrete example, manufacturing SMEs in the South Wales and Greater Manchester industrial areas of the UK are focused and their potential motivations for using and knowledge of big data-based customer analytics are investigated. An exploratory survey was conducted in terms of the type of customer data they have, the storage approaches, the volume of customer data, etc. Next, a carefully devised exploratory study was undertaken to understand how SMEs perceive the relations between customer data and product design, how about their expectations from big customer data analytics and what really challenges SMEs to exploit the value of big customer data. Besides, a demonstration platform is developed to present SMEs an automatic process of analysing customer online reviews and the capacity on customer insights acquisition and strategic decision making. Finally, findings from two focus groups indicate the different managerial and technical considerations required for SMEs considering implementing big data and customer analytics. This study encourages SMEs to welcome big customer data and suggests that a cloud-based approach may be the most appropriate way of giving access to big data analytics techniques.

Keywords:
Big data analytics; Online reviews; Customer insight; SME; Design innovation

1. INTRODUCTION

The manufacturing sector is undergoing a major IT and smart technologies enabled transformation, which provide on-demand computing services with high reliability, scalability and availability in a distributed environment and it is expected to become a major enabler for the manufacturing sector, transforming business models, helping it align product innovation with business strategy, and creating intelligent factory networks that encourage effective collaboration (Xu, 2012). The provision of decision support system services for the manufacturing sector through cloud computing enables the reduction in unit service costs due to the increase in operational size, the reduction in unit service costs due to the increase in number of services being developed and provided, and the reduction in unit costs due to the increase in number of services put through supply/demand chain (Demirkan & Delen, 2013). Besides, ICT and the internet have a similar transformational effect upon consumer retail not only from the perspective of pure retail, through online retailers such as Amazon or marketplaces such as Etsy, but also user-generated product reviews, provided through either retail websites or dedicated consumer review platforms such as Trustpilot. These have become increasingly important through the valuable information that can enable product designers to better understand the needs and preferences of consumers, whilst also influencing potential consumers in their purchase decision-making process (Y. Liu, Jin, Ji, Harding, & Fung, 2013).

In 2009, The Data Warehousing Institute (TDWI) conducted an extensive investigation of companies. Of the organisations surveyed, 38% reported that they utilise advanced data analytics, whereas 85% said they would be practising it during the next three years (Russom, 2009). According to the study, the respondents were spread evenly across a wide spectrum of company sizes. However, only 23% of respondents were from companies whose revenue was less than 100 million (Russom, 2009), within the EU this is above the €50million turnover threshold for a medium enterprise and most of the 23% of respondents are micro-enterprises with less than ten employees or small and medium-sized enterprises (SMEs) with 10 to 250 employees. Also, in a British government study conducted for the Department of Business Innovation and Skills reported that in 2015, 99.9% of the total number of enterprises operating within the UK can be classified as being SMEs. In total, these companies account for more than 2/3rds of the UK private sector workforce as well as 47% of the annual private sector turnover generated in the UK.

SMEs play an important role in the continued growth and success of national economies and studies have suggested that their contributions. Hence, the importance of SMEs should not be understated, in particular, their role in realizing the demands and improving the profitability of their supply chain partners, including large businesses, since that they have a critical role with a modern economy (Ghobadian & Gallear, 1996; Ward & Rhodes, 2014). Whilst there is much research regarding big data and collaborative design, there is comparatively little that examines these from the perspective of SMEs, despite their economic importance. Therefore, this work specifically focuses upon manufacturing SMEs, where is located in the South Wales and Greater Manchester industrial areas of the UK, and their potential motivations for using and knowledge of big data-based customer analytics.

In this study, the following two research questions are focused on.

1. What are manufacturing SMEs doing with regards to their product and customer data?

2. What position are SMEs in to make use of big data analytics?

The rest of this research is organized as follows. Section 2 discusses relevant studies regarding big data analytics as well as how big data analytics challenges SMEs and could facilitate SME design innovation. To investigate the aforementioned two research questions, in Section 3, the methodology is formally introduced, which includes how to design an exploratory survey and a case study, how to develop a demonstration platform, and how to organize an effective work focus group. Guidelines for designing a demonstrator system are presented in Section 4, and a prototype of the demonstration platform is presented in Section 5. In Section 6, key outcomes from the two SME focus groups are detailed. Finally, this study is concluded in Section 7.

2. BACKGROUND AND RELATED WORK

2.1 Big Data Analytics

The expression 'big data analytics' is a broad term that has given to the utilization of advanced data analytics techniques for the analysis of big data sets. Therefore it can be considered that big data analytics, which is often referred to as 'big data', is about two things. Firstly big data – the gigabytes or terabytes of data that a company can hold. Secondly analytics – the tools and techniques that are used to analyse the data.

In the 21st century, an enterprise will most likely have data that originates from many sources such as electronic business records, emails, production monitoring systems, social media, website logs etc. In the period up until 2003, a total of 5 exabytes (10^{18} bytes) of data were created by mankind, yet in the modern era, it takes only two days to create this volume of data (Sagiroglu & Sinanc, 2013). A study conducted by the McKinsey Global Institute (Manyika, et al., 2011) listed the potential benefits of big data analytics in five main domains, of which two can be considered as being relevant to manufacturing SMEs:

- Manufacturing: improved demand forecasting, supply chain planning, sales support, developed production operations, and web search-based applications.
- Retail: in-store behaviour analysis, variety and price optimization, product placement design, improve performance, labour inputs optimization, distribution and logistics optimization, and web-based markets.

As highlighted in (Mourtzis, Vlachou, & Milas, 2016) big data within manufacturing companies is not necessarily only about products and customers, a single machine tool monitoring system could easily generate nearly 0.5TB of data a year. Also, network-based control systems will generate potentially large quantities of data (Papacharalampopoulos, Stavridis, Stavropoulos, & Chryssolouris, 2016), with both of these sources containing potentially useful data. Chen and Zhang (Philip Chen & Zhang, 2014) stated that a large number of techniques and technologies to capture, curate, analyse and visualize big data have been developed by scientists in different fields such as computer science, economics, mathematics and statistics, but they are some distance away from being able to meet a wide variety of needs since that the field of big data analysis also encompasses data mining, machine learning, neural networks, social network analysis, signal processing, pattern recognition, optimization methods and visualization. Hence, the wide variety of tools, techniques and disciplines could make it difficult if not impossible for a non-IT manufacturing SMEs to implement without expert guidance and, as stated in (Philip Chen & Zhang, 2014), "Big Data also means big systems, big challenges and big profits, so more research works in these sub-fields are necessary to resolve it."

2.1.1 Customer Insights

Conventionally, customer insights can be obtained from formatted customer survey data and some researchers developed different models to identify customer insights. For instance, a hierarchical approach was introduced to solicit customer preferences from survey data (C.-H. Chen, Khoo, & Yan, 2002), which includes the verbatim construct, the superordinate construct and the imposed construct. Then, an ART2 neural network was built to analyse the customers. The results from the neutral network help to analyse customer segmentation and to conduct the market analysis. However, consumers might present different interests in product aspects. To cluster consumers with similar interests, a permutation-based structural topic model was proposed (Si, Li, Qian, & Deng, 2014). Using this model, the frequency of different product aspects and the occurrence order are presented.

Compared with conventional customer survey data, a large volume of online data is available online, which could enable product designers to understand consumer insights with a finer granularity. For instance, the search information about products from Google trends was utilized to analyse consumers' attitude (Jun, Park, & Yeom, 2014). This search information is argued to be better in obtaining consumers' level of interest in the product level and comprehending the product features they concerned as well as their importance. Also, this information is said to be helpful in the prediction of consumers' preferences. Another typical scenario is that consumers might visit different online shopping stores to make price comparisons of the same product. To facilitate price comparing, a rule-based personalized comparison method was proposed (G. G. Lim, Kang, Lee, & Lee, 2011). Some typical price comparison sites neglect the precise delivery cost. In this rule-based method, shipping rules, destination, delivery speed as well as shipping ratings are taken into consideration. Consumers often have to make decisions with alternatives in online shopping and various product comparison agents are built for this purpose. Product comparison is conducted from both product-based and consumer-based. Accordingly, Wan et al. categorized product comparison agents into three types: product differentiation, product evaluation and consumer preference identification (Wan, Menon, & Ramaprasad, 2007). The first type focuses on objective information. Product evaluation type is on subjective information such as services or the food quality of a restaurant. Consumer preference identification type takes consumer subjective textual feedbacks into consideration. Also, some product comparison agents involve information from three different types of agents, in which obtain evaluation information, differentiation data and user experiences.

2.1.2 Online Consumer Data

The rapid growth of the internet, when combined with the rise of ‘web 2.0’ i.e. user-generated content such as internet forums, social media and review sites, has created a large volume of content that is of potential interest to researchers. However, there is one significant difference between the data gathered by legacy systems from the 1980s and ’90s and that of web-based and e-commerce systems. Namely, the data is less structured and often contains rich customer opinion and behavioural information, in particular, user-generated reviews at e-commerce sites for example (H. Chen, Chiang, & Storey, 2012). Such rich information could be of particular use to the manufacturers of consumer goods, in particular, SMEs who may not be in the position of large companies who are able to conduct extensive market research.

The effect of online reviews on product experience is analysed in (Hu, Liu, & Zhang, 2008). They found that samples of high-quality products can reduce product uncertainty. Also, the impacts of online reviews on sales is lower, comparing with experience products with samples than without. Furthermore, research suggests that people rated positive reviews as more helpful than negative reviews, whilst it took longer to evaluate negative reviews (Guo, Zhao, Zhang, & Wen, 2019). In addition, the uncertainty that is induced by online reviews can be softened by providing samples. Also, the demographics of customers was learned from purchased items, which was formulated as a multi-task multi-class problem (Oentaryo, et al., 2016). Specifically, a structured neural embedding model was proposed to learn the representations of items and the correlation information between different tasks was explicitly modelled by turning the multiple multi-class prediction tasks. Similarly, a logistic regression-based framework was proposed to predict the label of users in social media (Wang et al., 2016). Specifically, a semi-supervised learning approach was introduced, in which multiple relations were reckoned. Then, the objective was modelled to obtain a classification model such that the expectation of its predictions on unlabelled data is similar to the class prior.

Besides, some research studies intend to extract customer insights from customer online data, such as customer reviews, tweets, etc. A Bayesian Sampling approach was proposed to extract product features from online reviews (S. Lim & Tucker, 2016), in which the term disambiguation problem and the keyword recognition problem were investigated. First, a set of optimal search keywords are determined to avoid the false positive of the identification of product feature related sentences. Then, the sampling starts from data that involve a product feature related item and tries to maximize the F-measure about both two problems iteratively. The product favourability was defined as the product of polarity, subjectivity and popularity (Tuarob and Tucker, 2015). The favourability was then utilized to select the most and the least favourable products. Next, the approach of LDA was applied to identify product features from tweets. Some researcher also proposed a method to identify opinion leaders in a specific domain (Miao, et al., 2013). Customers post several reviews and reviews may belong to the different domain and, in this method, the number of reviews in the same domain is utilized to define the similarity of consumers. The integration of topic, sentiment, and syntax for the purposes of modelling online reviews has been investigated (Tang, Jin, Liu, Li, & Zhang, 2019). A probabilistic topic model termed tag sentiment aspect models were proposed on the basis of Latent Dirichlet allocation to reveal latent aspects and corresponding sentiment in a review simultaneously. The extraction of user experience data pertaining to the underlying product design from one customer online reviews has been examined (Yang, Liu, Liang, & Tang, 2019). Which uses user experience discovery to extract data from single reviews, followed by data integration to group similar data and final network formalization to build up the causal dependencies among user experience groups.

2.2 Challenges of SME Big Data Analytics

Whilst it is incorrect to assume that all SMEs cannot afford sophisticated computer systems, cost and in particular value for money often play a critical role and hence potential barrier when a company is considering making a software purchase. There is a natural cost – functionality trade-off, systems may have functionality that is desirable but the additional costs outweigh any of the potential benefits. An equally important consideration is that SMEs often do not have in-house capabilities for the selection, installation, configuration and maintenance of complex IT systems - such factors are a potential barrier to big data analytics within SMEs (Schmidt & Möhring, 2013). The main challenge related to the use of Big Data, specifically the skills for handling it, has been identified as being of particular concern for SMEs, as not only are the skills difficult to find but they are, most importantly, expensive to acquire (Del Vecchio, Di Minin, Petruzzelli, Panniello, & Pirri, 2018).

Indeed the problems for SMEs (whether manufacturing or not) with regards to big data analytics have been shown to be complex and multifaceted covering factors such as IT, data analytic intelligence, organisational structure, data security, and legal aspects (Coleman et al., 2016). With Coleman et al. stating that "What is now needed is a mechanism to help SMEs to get started".

The emergence of cloud computing and Software as a Service (SaaS), could provide a means to access complex databases and IT systems. Studies have suggested that cloud computing will become an attractive option for many SMEs due to its flexible cost structure and scalability (Sultan, 2011). Although interoperability between cloud computing platforms and applications as well as cloud and 'desktop' systems may present challenges (Dillon, Wu, & Chang, 2010).

As stated in (Schmidt & Möhring, 2013) the application of cloud computing and SaaS technologies to big data analytics has the potential to be very cost-effective. In particular for

SMEs where cloud computing can reduce the cost of implementing big data analytics and make it feasible for them to use it. Despite the positivity associated with the concept of big data analytics being provided as a cloud-based service, there are still a considerable number of challenges technological and organisational that need to be addressed (Ardagna, Ceravolo, & Damiani, 2016).

2.3 Design Innovation in SMEs

With the fast development of Web 2.0, a large volume of online social data presents instructive customer expectation, which is definitely beneficial for SMEs who usually do not have a capacity to obtain sufficient customer feedbacks dynamically. To promote competitive and innovative products according to customer concerns, a Customer Relationship Management (CRM) system is outlined through mining the value of social media (Ajmera, et al., 2013). It is argued that a social CRM system needs to process informal and heterogeneous content, identify relevant and actionable posts, determine the priority of posts and deal with the bootstrapping challenges. Another system that monitors customer online opinions was built (Goorha & Ungar, 2010). In this system, frequent phrases and phrases near the terms of interest were initially extracted from textual data, which were utilized to identify which of them appear dramatically. Then, whether a phrase is interesting was depended on how often they are referred to, how often they are referred comparing with before and how specific they refer to a topic.

As one typical resource of customer opinions, online reviews are utilized for product selection and trend prediction of design innovation. For instance, a three steps method for customer-driven product design selection is proposed by analysing customer online reviews (Wang, Xie, Liu, & Philip, 2011). In the first step, product attributes were extracted from online reviews. In the second step, a hierarchical customer preference model was built by using

Bayesian linear regression. Product ratings, category ratings, attribute ratings and product specifications are taken into consideration in this hierarchical model. Finally, an optimization problem was formulated to maximize potential profit by taking engineering constraints into consideration. Tucker et al. employed online reviews to predict trends of product design innovations (Tucker & Kim, 2011). Sentiment polarity of product features was extracted from online reviews. Then, the Holt-Winters exponential smoothing method was employed to model product preference trends.

Also, another widely utilized approach to derive customer insights for design innovations is to make effective comparisons with competitors since consumers often make decisions with alternatives in online shopping and both product-based and consumer-based comparison perspectives are claimed to facilitate product comparisons or product ranking. For instance, Wan et al. categorized product comparison into three types, i.e., product differentiation, product evaluation and consumer preference identification (Wan, et al., 2007). Product differentiation focuses on objective information, product evaluation is on subjective information such as services or quality and consumer preference identification type takes consumer subjective textual feedbacks into considerations. A graph propagation method was proposed to compare products in considerations of online reviews and community-based question answering (Li & Zhan, 2011), in which comparative sentences are extracted from online reviews and the information about the number of preference between two products is utilized in the graph method. A product comparison network was reported by exploiting comparative sentences in online reviews(Zhang, Guo, & Goes, 2013). Three different types of graph are built according to the number of the overall sentiments of consumers, single-link graph, dichotomic-link graph and multi-link graph. Also, different regression models were utilized to analyse the factors that influence the product rank. According to Zhang, et al., 2013, a new product comparison network was built and, rather than the averaged sentiment, transitive influence mechanisms, denoted as transitive sentiment, were utilized to analyse the network influence (K. Chen, Luo, & Wang, 2017). Next, PageRank and HITS centrality, average rating, average sentiment, transitive sentiment, category dummies and time dummies were combined into a linear model to analyse the importance of each variable and predict sales rank of four categories of products.

3. METHODOLOGY

The primary data gathering aspect of this work takes primarily an inductive approach and it encourages the use of research instruments that are capable of delivering principally qualitative data. The use of qualitative research is suitable in situations where the aims are to understand the meaning that participants in the research give to events and situations, and to appreciate the context in which these are made (Maxwell, 2012).

3.1 Managerial Principles of Exploratory Survey

In order to inform the work, a total of forty online surveys were completed by participants who have senior positions in Welsh manufacturing companies and a broad spectrum of manufacturing sectors were surveyed. This online survey incorporated aspects of semi-structured interviews by allowing the respondents to enter free form text. This was used to try and achieve higher levels of participation from those in managerial positions, to allow participants to understand how the information they provide will be used in terms of addressing trust issues, and to negate some of the effort associated with writing (Lewis, Thornhill, & Saunders, 2007). A coding schema for this survey was developed, and the results were used as a contributor to the research objectives of this study and for the framing of the subsequent studies.

The demonstrator is intended to expand upon the exploratory survey, helping to develop theory and further inform the work to be undertaken during the prototype development and focus group phases. Literature states that cases should be purposefully selected to best illuminate the phenomena that are being scrutinised (Yin, 2003) and it is considered to be important that case studies should be selected with a sense of purpose (Stake, 1995). The post-demonstrator focus groups would enable feedback to be obtained from relevant parties.

3.2 Design of a Demonstration Platform

The goal of the demonstration platform is to automate the process of analysing customer online reviews and to develop a scalable, user-friendly system to obtain customer insights that inform corporate strategic decision making. The system will gather data from websites, categorise the review texts into a sentiment polarity, i.e., positive, negative or neutral, and provide product summarization based on features of a given product in real-time. Essentially, it forms a data ingest pipeline from data collection, analysis, interpretation and visualization. The primary objectives of the system design include the following four aspects.

- To facilitate the search for a product and the associated review data.
- To acquire the review data and perform sentiment analysis on it to extract useful features and their sentiment.
- To interpret and visualize the analysis results in a meaningful manner.
- To provide an efficient and user-friendly web-based interface to control the system.

Based upon the system objectives listed above, the technology demonstrator needs to be able to provide the ability for a user to search for products, select them and view charts that highlight the customers' sentiments on features the product has. In addition, more customer sights should be covered, such as, the overall product satisfaction, a list of features specifically chosen for comment by customers and the associated sentiment, a comparison of categories for a given polarity to see sentiment distribution, the distribution of sentiment for any given category, and the comparison of these charts for different products. In addition to those features, the system should illustrate the potential for geographical visualisation, highlighting where the products were liked or disliked or areas with a large number of reviewers or particular sentiments. Besides, all of the review text should be accessible allowing the user to select any section of the visualisations and read the appropriate reviews. Thus allowing the user to read why there is a particular sentiment regarding the product and/or its features. Such information can be beneficial for product designers to help improve products or services by highlighting problems areas or showing what works well (Y. Liu, et al., 2013).

3.3 Workshop Focus Groups

The utilisation of focus groups can be helpful in the process of identification of major themes and are considered to also be useful for exploratory investigation of particular issues (Krueger & Casey, 2009). The optimum number of focus group participants is considered to be between six and ten, and that the composition of the group should be balanced with regards to similarities and differences in participants (Krueger & Casey, 2009). As such, the focus groups involved participants from micro to medium-sized enterprises in addition to academics and knowledge transfer practitioners, plus a facilitator and rapporteur.

Two focus group activities took place as part of workshops organised in Cardiff and another in Manchester, with a total of 30 participants. To stimulate the discussions, the participants were presented with the outcomes of the survey and case study, plus the customer sentiment analysis demonstration platform. Notes of the outcomes of sessions were taken and formally compiled to identify the key themes and challenges raised.

4. GUIDELINES FOR DEMONSTRATOR SYSTEM

The demonstrator is heavily informed by the responses of the wide variety of companies surveyed. However, the application domain is particularly focused upon the needs of companies manufacturing goods that are retailed through third parties. To create a picture of the challenges that SME faces, survey results have been synthesised. Illustrating what an exemplar of a consumer goods manufacturing SME is doing, wishes to do and what challenges they face.

4.1 The Exploratory Study of SME Data Use

Typically firms will possess data related to company sales and orders, and some have product information. As Table 1 shows, the vast majority of survey respondents have product data and that sales data is maintained by the majority. With the benefit of hindsight, the near-universal maintenance of product data should be considered to be unsurprising as it the scope of the definition is sufficiently broad that it could incorporate CAD data, product specifications, product datasheets etc. Despite companies appearing to maintain data, Table 1 does not indicate how much data is actually related to customers. When asked if they maintain data about their customers, it was found that 80% said that they do and 20% that they do not.

Table 1. Types of data held by companies.

Type of data held	Percentage of respondents
Supplier Data	77%
Manufacturing Data	77%
Product Data	91%
Sales Data	82%
Customer Data	68%
Financial Data	77%
Other	14%

Respondents that maintain customer data were requested to provide more information regarding how they currently use data, irrespective of how it is stored, the results are shown in Table 2. Of those that provided a response, the majority indicated that they use it to improve their products, market and customer service. As some of these are provided within CRM systems respondents were asked to state whether they used CRM - it was found that 67% of the respondents did not use CRM. This raised a further question regarding how many companies use or keep records of their product reviews, discussions at the SME focus groups suggested that although reviews might be read there is no systematic evaluation of the reviews.

Table 2. Current uses of customer data.

Use of data	Percentage of respondents
Personalising Customer Experience	31%
Improve Products of Services	69%
Improve Marketing	75%
Improve Customer Service	75%
Reduce Risk and Fraud	19%
Other	13%

Studies have suggested that cloud computing could be an enabler for SMEs that desire to utilise big data analytics (Schmidt and Möhring, 2013). However, to make this achievable, there will be a necessity to make some use of cloud storage. Therefore the respondents were asked what storage methods they made use of as presented in Table 3. It shows that 50% of respondents use cloud storage for some if not all of their data, a broadly positive result as it implies there is a degree of familiarity with cloud computing at some level.

Table 3. Storage methods used by SMEs.

Storage type	Percentage of respondents
Flash Drives	27%
Hard Drives	55%
Hard Drive Arrays	23%
Network Attached Storage	64%
Cloud Storage	50%

To gauge how much data manufacturing SMEs actually have, the respondents were asked to estimate how much data their company actually has, as shown in Figure 3. The majority of the respondents' estimate they have less than 2TB of data. Within the context of big data, this amount may seem comparatively small. However, when a historical perspective is taken, 500GB of data is over 100 times the capacity of a typical hard drive used in a fileserver's RAID array in 1996 (c. 4.3GB) and as such would have been unimaginable for SMEs a little over 20 years ago. When asked if they were aware of big data analytics only half of the SMEs said they were aware of big data analytics. A figure of 85% was given by TDWI for companies that would be using it within three years (Russom, 2009). Suggesting a potential mismatch between what businesses as a whole and what manufacturing SMEs think about big data and also their ability to gain meaningful results. Of particular note is that only 25% of the companies surveyed by TDWI were SMEs and that only 4% were non-IT manufacturers, suggesting that larger enterprises are more aware of big data than SMEs.

Table 4. Amount of data stored by SMEs.

Volume of data	Percentage of respondents
1Gb – 500Gb	24%
500Gb – 2Tb	33%
2Tb – 10Tb	29%
>10Tb	14%

The SMEs were then asked what sort of information they would hope to garner from the application of big data analysis and results are shown in Table 5. When compared against the lists provided in (Manyika et al., 2011), it could be concluded that the overall awareness of what big data analytics could achieve is perhaps not as good as it could be. Furthermore, it suggests that companies may have heard of big data analytics but are, in reality, unaware of its capabilities. Once a cross-tabulated with the result about the awareness of big data, three-quarters of the respondents who had heard of big data analytics actually had a vision of what they'd hope to achieve through its application. During the focus groups the comment was made that "even big SMEs struggle with Little Data, let alone Big" which participants broadly agreed with.

Table 5. Insights that companies wish to gain through analytics.

Insights companies wish to gain	Percentage of respondents
Market Understanding	27%
Increase in Productivity	27%
Understanding Customer Patterns and Characteristics	36%
Sales Predictions	36%
Other	46%

Based on these initial survey findings, and the discussions with SMEs it can be concluded that there is a requirement to further inform SMEs about big data analytics, what it can be used for and how to overcome barriers, real or imaginary, to entry that SMEs might face.

4.2 Customer Engagement and Collaboration

Amongst the SMEs surveyed, it was discovered that there were essentially three main modes of engagement. Direct sales from the manufacturer to either a retailer or wholesaler, retailer sales to customers via a physical and/or internet retailer, and retailer sales via internet-based marketplaces such as eBay or Amazon marketplace.

The information flow is shown in Figure 1. As can be seen, there can be multiple barriers in getting feedback to a manufacturer. In certain retail situations, the information flow may potentially be indirect and will require effort to obtain a review of a product, especially when feedback is left on internet review sites. Companies, by having no direct or indirect contact with their customers, dealing solely with resellers through salespeople, reported that they find it difficult to obtain information. This can include salespeople, working for the company, filtering out "bad news" that they think the manufacturing operation does not want to hear, even though they do want this information to improve product quality. It was specifically noted that the companies in question do not directly search internet retailers or review site to obtain customer feedback.

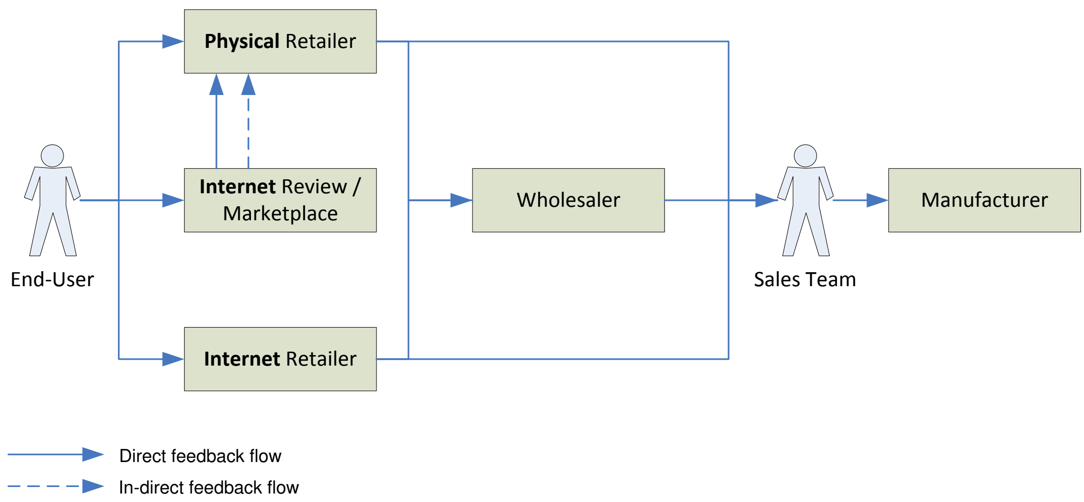


Figure 1. Customer feedback information flow.

As well as the challenges associated with obtaining feedback, another issue that companies face is that of regulations. At one level it can be CE marking which indicates that the manufacturer claims compliance with the relevant EU legislation. However, a far more significant challenge is faced by companies questioned that have far tougher regulation regimes because of the sector they operate in and/or the products they manufacture. As an example, some safety products require type-approval, as such they are produced to a standard with little room for manoeuvre before re-approval is required. So, for example, colour can be changed to meet consumer demand but if the geometry or materials changed then new type-approval is required. This can tie the hands of the manufacturer and make customisation prohibitively expensive.

4.3 Customer and Product Data

The companies questioned can maintain a variety of data from and regarding their customers. As mentioned, there are several different sales channels that the companies can use. The data is collected from their sales team and would come from the retail channels that they serve. One company conceded that they do not get as much information as they would like and moved their managing director from the manufacturing location to be closer to the sales team to get better feedback.

The specific product data can range from the basic data relating to the design of the product, to more detailed and specific data relating to standards compliance. The availability of product feedback data can suffer from sales teams sometimes filter information such that product designers do not get a full picture of what is going on. It makes the availability of some form of analysis tool where customer options and preferences could be gathered and analysed a highly attractive proposition.

Challenges include not having as much feedback data as they would like and data and IP protection, specifically stopping competitors copying products.

4.4 Big Data Enabled Product and Service Design

Although companies had some awareness and interest in the concept of big data, there was no real knowledge of where to start when it came to implementing it themselves. Especially as much of the discussion seems to be centred on topics such as extracting information from supermarket loyalty card data for example.

Amongst the things companies are interested in finding out via big data are developing trends within the market that they serve, feedback on the products they manufacture and what gaps in the market could be based. Based on this data, their products could be improved and new ones potentially developed.

5. DEMONSTRATION SYSTEM

As shown through the survey of companies only 50% of SMEs are familiar with the concept of big data and even those that are aware of it may not fully understand its capabilities. This confirmed the necessity of providing a demonstrator platform that shows SMEs can better understand how customer analytics and big data analysis could be of benefit to them. The overview of the system is shown in Figure 2.

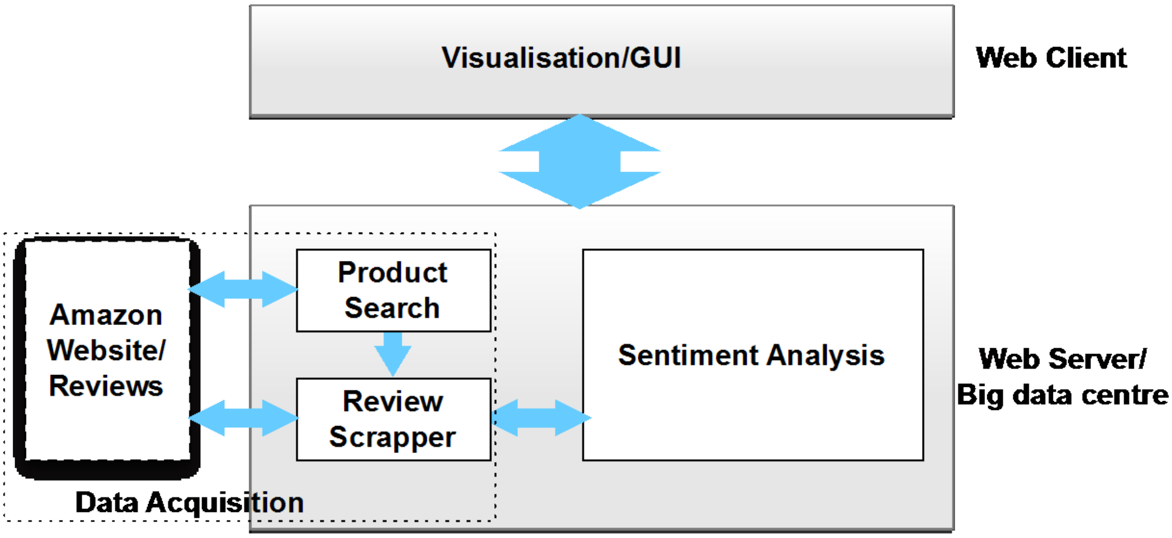


Figure 2. System overview

Because of the variety of products sold and the volume of product reviews available, Amazon.com was chosen for the purposes of the demonstration system. However, the system can be adapted to use content from other websites for sentiment analysis of product reviews. There are four main components of the demonstration system, including Data Acquisition, Sentiment Analysis, Data Visualization and Graphical User Interface.

5.1 Data Acquisition

To analyse product reviews, it is important to first accurately identify the reviews and collect data from the websites in a timely manner. However, due to the high volumes of data involved, the manual collection of web data has a significant labour cost burden and tends to induce "human error". As such there is a requirement to develop an automated approach for review website data scraping. In this study, a web scraping service was developed to collect review data from the Amazon website, which can be utilised by the analysis components of the system. Due to restricted access to the review information in its raw form, the Amazon product API was combined with the web scraping service to search the Amazon database and return the search results. This then prompts the automatic retrieval of the review data and loading of the product review pages from the website and acquisition of the text of reviews for the subsequent stage of analysis.

5.2 Sentiment Analysis

Sentiment analysis refers to detection and classification of whether a given text represents a positive or negative or neutral opinion. There are three levels of sentiment polarity classification including document, sentence and aspect levels (B. Liu, 2012). The document level mainly concerns with classification of an entire document as expressing either a positive or negative opinion. The sentence-level analyses the sentiment polarity of each sentence. The aspect-level sentiment analysis primarily deals with specific aspects or features of a document. The demonstrator system uses aspect-level sentiment analysis of product features. For example, a mobile phone review might be, "The screen was great but the speakers were terrible". An ideal sentiment analysis system ought to be able to identify both "screen" and "speakers" as product features with positive and negative sentiments respectively. Aspect-level sentiment analysis is particularly useful when comparing two or more products based on their feature aspects. However, most current websites only provide an overall score for a particular product. Therefore consumers have to read all of the reviews and conduct a manual analysis in order to make their decision.

5.2.1 Workflow

In the demonstrator system, both document-level analysis and aspect-level analysis have been applied to product reviews. The workflow of sentiment analysis module is shown in Figure 3.

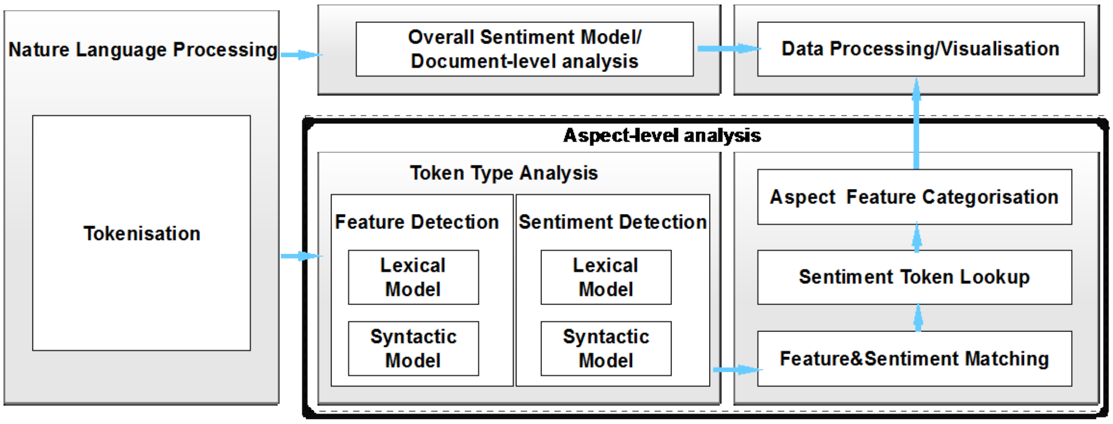


Figure 3. Sentiment analysis workflow.

As shown in Figure 3, the Natural Language Processing (NLP) module takes review text as its input and pre-processes the data by performing tokenisation on the texts for further aspect-level and document level analyses. The aspect analysis module receives its input from the NLP module and identifies features and sentiments as well as performing aspect feature categorisation. This can then be visualised by the user through the graphical user interface. Similarly, the document level analysis performs sentiment analysis of the whole review text with the results being displayed to the user.

5.2.2 Aspect-level and Document-level Sentiment Analysis

For the aspect-level sentiment analysis of the demonstrator system, the work of Carter (Carter, 2015) was built upon and extended by introducing a new feature categorisation and a feature matching algorithm. The core technologies utilised are the Stanford Natural Language Processing library (coreNLP) for tokenisation (CoreNLP, 2016; Manning, et al., 2014) and Support Vector Machine (SVM) using the libSVM library implementation for feature detection and sentiment detection (Chang & Lin, 2011).

Essentially, the system attempts to determine if any given word within a given review is a 'feature' or a sentiment-bearing word by first processing the review text using coreNLP which tokenises the text and assigns features to each token based on natural language principles. They are then passed into the four SVM models, including both lexical and syntactic models for 'feature' and 'sentiment' detection separately. These models then provide a score based on how much each token fits each role and if a token has scored above zero then the model that provided the highest score determines its type. Once the features and sentiments have been extracted, it determines which sentiment applies to which feature through the use of the matching algorithm developed for this demonstrator. This data can then be collected such that all the raw review texts are packaged up with their specific extracted features and sentiments and sent to the web application.

An important part of this analysis concerns the matching of each feature with the appropriate sentiment bearing word so that the correct sentiment value can be recorded. A matching algorithm has been developed that follows a rule-based approach to connecting a feature with its appropriate sentiment working first on a proximity basis and then to any connected sentiment found through the natural language processing. This method consists of three steps:

Step One check whether the feature itself is sentiment bearing and if so matches the token with itself as the sentiment expressed by the token is certainly related to the feature, i.e. "comfortable".

Step Two iterates through the surrounding six tokens, three on each side of the feature looking for a sentiment. In this case, if the sentiment is found, it is matched because the closer to the feature the sentiment is, the more likely it is to be related. For example, many of the reviews had simple terms such as, 'good looking' or 'great price' as well as 'the colour was good' or 'it fits very well' and as such by checking the six tokens alternately starting with the preceding one and then the next one and continuing outwards any detected sentiment had a high probability of being connected. However a number of issues can arise from this particularly lists and comments including 'but'.

Step Three checks the semantic incoming edge of the feature token, which is set as part of the natural language processing module, in order to determine whether it is a sentiment or a feature. If it is a sentiment then as with the previous reasoning it is likely to be connected to the feature as it is in semantic proximity. If it is a feature then the process is that given that the feature is semantically close and a sentiment has not been discovered within a three word radius, then the sentiment of the new feature is assigned to the current one as it is likely with a list or comment discussing the same sentiment about multiple features at the same time. To do this the matching algorithm is recursively called on the new feature and the result returned. Finally, the other semantically tagged tokens are inspected.

For the document-level analysis, the sentiment polarity of a whole review text has been analysed utilising coreNLP, which takes the input of a review text and returns the polarity of each sentence and then aggregates polarity of sentences.

5.3 System Implementation

To enable the system to be run on the cloud computing platform and ensure scalability and efficiency, a REST design pattern was chosen. All the system components have been implemented as web services based on Dropwizard, a java-based framework/library to facilitate the development of high performance, restful web services. The system combines Jetty as an embedded web server, with Jersey to map Java objects directly to HTTP requests using Jackson for the JSON conversion and a variety of often used libraries in Java development. Essentially, the creation of standard Java classes and the use of annotations to map the various methods and object to the HTTP request URLs without having to manually create handlers or servlets, etc. It also generates a single large jar file which combined with a configuration file is all that is needed to run the service, making distribution and deployment a simple matter.

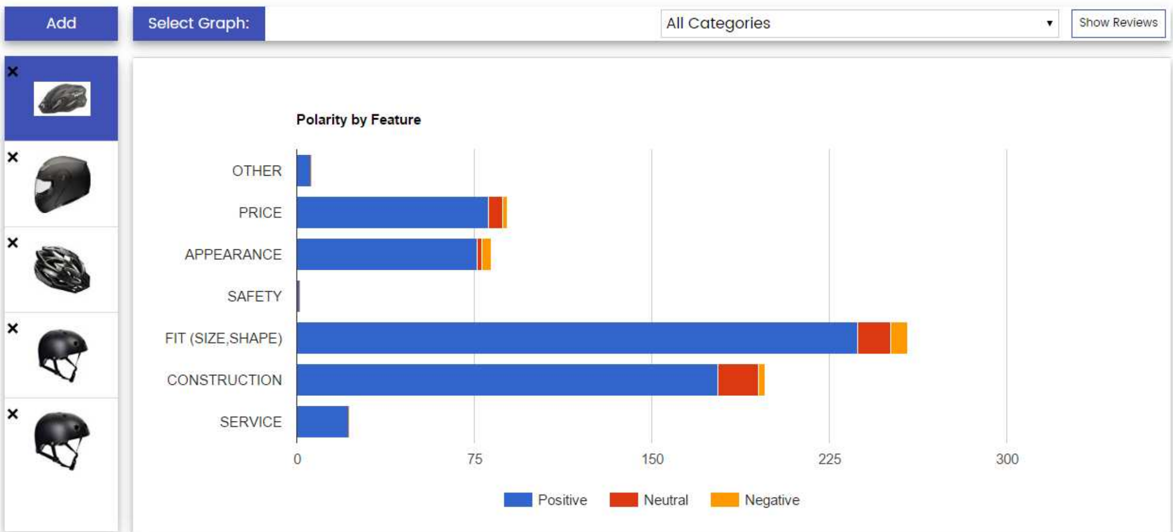
5.3.1 Data Sources

The data was scraped from the Amazon website in real-time. Since a supervised machine learning approach was utilised for training purposes, an initial sample set of 704 reviews containing 1,073 product aspect features was collected for the model construction. Each feature was tagged with the exact text from the review that represented. N-fold cross-validation was used for the validation of the model.

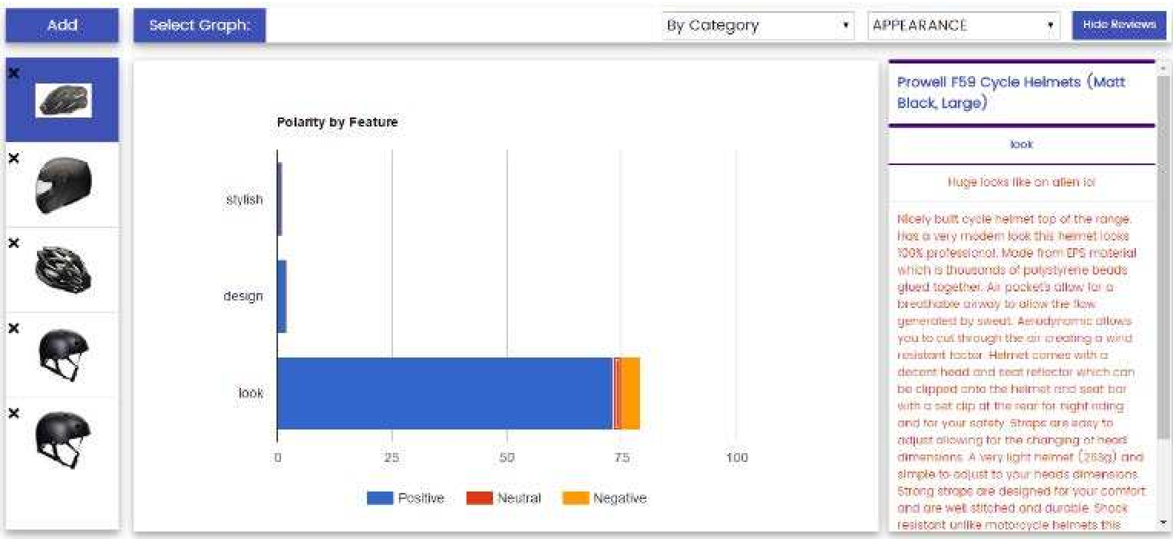
5.3.2 Data visualization

To provide a clearly condensed representation of the reviews about a given product that a user wants to get, a variety of visualization techniques have been applied to create charts that provide product review summaries, which are based on features and or aggregated features of a given product. The charts have been chosen to provide insights based on the results of the analysis but also to show the various different ways the data can be displayed. These charts include:

- A stacked bar chart that shows the positive, neutral and negative reviews for all features, as shown in Figure 4 a) and b). Through this chart, the end-user can immediately understand how customers feel about specific aspects of a product. From Figure 4 a) it is easy to see that most customers care about "Fit" and "Construction" with the majority of reviews being positive.
- A pie chart is utilised to show multiple products with positive, neutral and negative for each feature, as shown in Figure 5. This chart allows the end-users to quickly determine which products, categories and features have strong positive or negative opinions.
- A column chart is employed to show multiple products with positive, neutral and negative for each feature, thus enabling the comparison of product features, as shown in Figure 6.
- A geo-location chart showing positive, neutral and negative reviews by country of manufacturer, to allow geographic traits to be identified as shown in Figure 7.



(a) Polarity by category without showing review texts



(b) Polarity by category with review texts

Figure 4. A stacked chart to provide summarization of polarity for each category of a given product

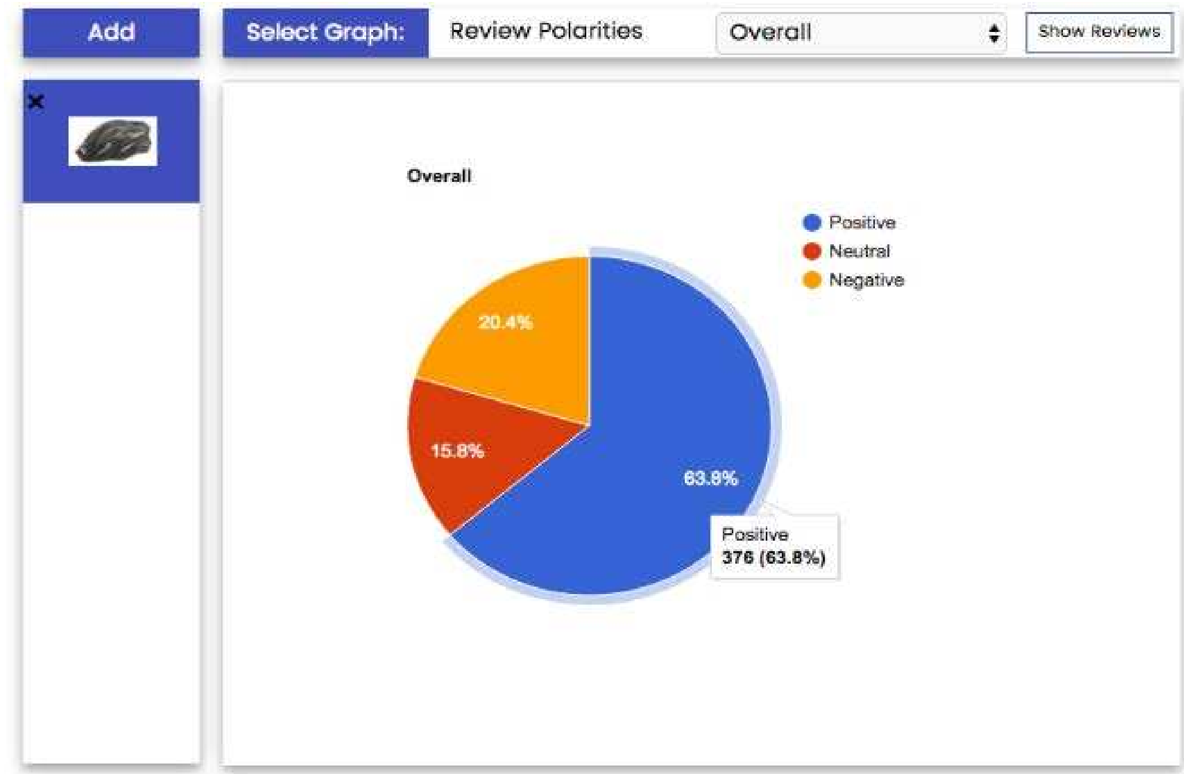


Figure 5. A pie chart to summarise the polarity of reviews.

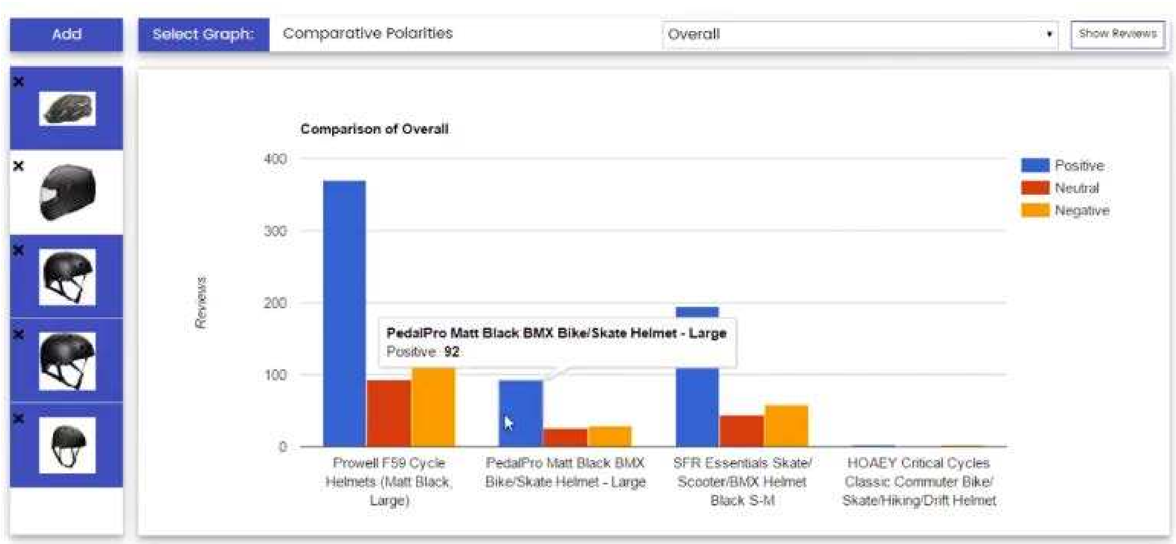


Figure 6. A column chart to show multiple product features for easy comparison.

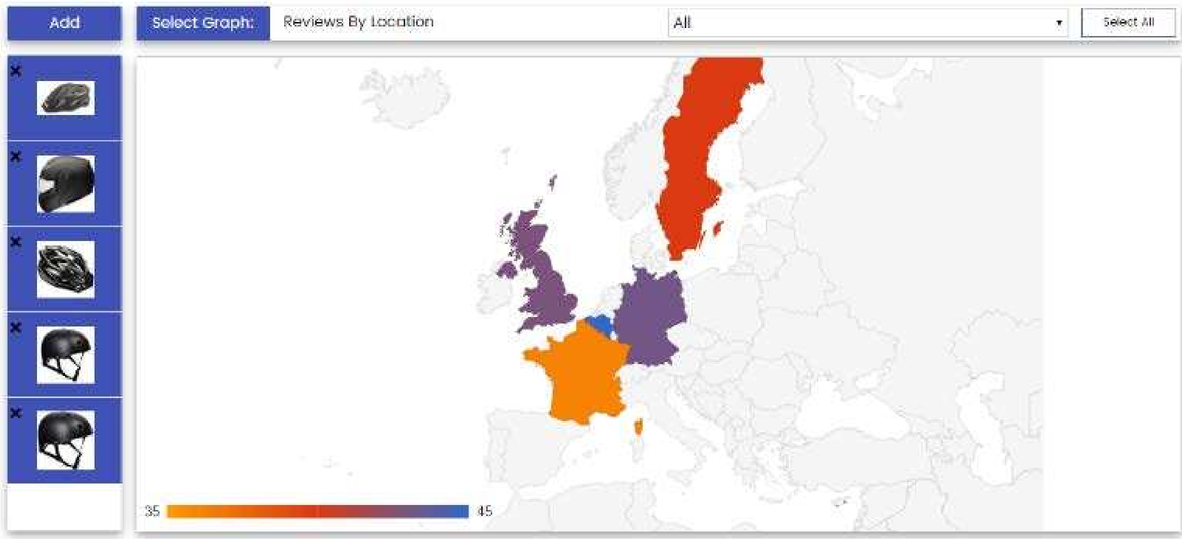


Figure 7. A geo chart to show positive, neutral and negative reviews by country of manufacturers (The colour of the country represents a normalised polarity).

6. FOCUS GROUPS

Two workshops in Cardiff and Manchester were organised to present the outputs of the SoC-UDRdM project, in total there were some 30 attendees from business and academia including facilitators and rapporteurs. The participants were presented with the results of the survey and the data analytics demonstrator. These presentations were used as an instigator for the focus group discussions.

The common theme arising from both of the workshops was that there was a large knowledge and capabilities barrier preventing SMEs from even beginning to utilise big data analytics, let alone maximising the potential benefits to improve their products and customer service. The key points that provide potential managerial and technical implications arising from the workshop focus groups are provided in the following subsections.

6.1 Big Data Demonstrator Platform

The general view of the demonstrator amongst the focus group was positive, it was felt that the demonstrator was a good starting point to achieve the type of application an SME would be looking for. The idea of having a centrally maintained (cloud) based tool was of particular interest as it would remove some of the barriers to entry. It was noted that social media is becoming an increasingly more important channel for feedback from customers (both B2B and B2C clients). Because of this, the need for sentiment analysis will grow together with that for tools which SMEs can easily access and use. Furthermore such a tool would help SMEs overcome two of the main barriers to uptake of new technology: cost and lack of know-how.

6.2 Managerial and Technical Implications of Big Data

"Little Data" might be a more appropriate term for the application of Big Data analysis techniques to SME data. It is also the case that even large SMEs can struggle with "Little Data" let alone "Big". Also, Big Data is context-sensitive e.g. an SME that uses computational modelling and CAD tools will have considerably more data than one that does not, yet relatively speaking they both have big amounts of data. Fundamentally SMEs (as well as enterprises in general) need to know: What they can do, Where the data is, how they can analyse the data before they are in any sort of position to begin applying the techniques. This application of Big Data, plus the Internet of Things, could then act as a precursor for moving towards a product-service system model of operating.

SMEs are concerned about choosing the wrong tools and techniques for analysing because they have insufficient knowledge regarding Big Data. Cost is a concern and SMEs are not always in the position to be able to spend large amounts of money on Big Data tools, even if they did they would not necessarily have the human resources to configure, run and maintain such a system. Cloud-based tools that are low cost with access to sources of product review data such as Amazon or other internet retailers appear to be desirable as it would help overcome the skills-gap hurdle that SMEs will face

6.3 Managerial and Technical Implications of Customer Insights

Customer insights are not guaranteed to deliver the "next big thing", as customers have a tendency to be reactive rather than proactive, although they may deliver product improvements. Additionally, insights needed into how customers use a product rather than just what they want or think of a product. This is because sometimes what customers think they need from a product can actually be different from what they actually need and Big Data could help find this out.

Social media is becoming increasingly important for feedback, both good and bad, for both B2C and B2B customers. Therefore through tapping into this information, the customer insight system seems to be a good starting point to achieve the type of application that an SME would be looking for.

7. CONCLUSIONS

This study has investigated the use of big data tools in product design and improvement from the perspective of SMEs.

The survey and focus group have shown that SMEs have product and customer data which the majority use in one way or another to improve their products and service. However, they are not necessarily utilizing what they already have in a truly effective manner as 67% of the respondents did not even use CRM. This could reasonably be considered to be a matter of some concern because if a firm is unwilling or unable to use CRM a reasonable question can be raised regarding their ability to fully utilise the benefits associated with big data analytics.

Even though SMEs are in fact aware of big data (albeit limited) they do not know where to begin. They feel that the term Big Data is overwhelming and does not reflect the reality of SMEs and that "Little Data" might be a more appropriate term for SMEs and that what a large volume of data is will vary from sector to sector. The work has identified that SMEs are particularly concerned about choosing the wrong tools and techniques for analysing data because they have insufficient knowledge about the subject and do not necessarily have the time to invest in investigating different solutions. As well as the fear of not selecting the solution that becomes the industry or market standard and that they have to spend money on changing systems. As previously mentioned when presented with the prototype system, they were very positive about the concept and could begin to see how Big Data could be of use for them.

Overall, the idea of Big Data analytics for gaining customer insights was well received but SMEs will need both guidance and the suitable tools before they will be able to make the most of it. The feedback strongly suggested that a cloud-based approach may be the most appropriate way of giving access to big data analytics techniques to SMEs. This concurs with the findings of Sultan (2011) that stated that cloud computing is an attractive proposition for many SMEs.

Acknowledgements

To be inserted...

REFERENCES

Ajmera, J., Ahn, H.-i., Nagarajan, M., Verma, A., Contractor, D., Dill, S., & Denesuk, M. (2013). A CRM system for social media: challenges and experiences. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 49-58): ACM.

Ardagna, C. A., Ceravolo, P., & Damiani, E. (2016). Big data analytics as-a-service: Issues and challenges. Paper presented at the 2016 IEEE International Conference on Big Data (Big Data).

Carter, D. (2015). *Inferring Aspect-Specific Opinion Structure in Product Reviews.* Université d'Ottawa/University of Ottawa.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST), 2*, 27.

Chen, C.-H., Khoo, L. P., & Yan, W. (2002). A strategy for acquiring customer requirement patterns using laddering technique and ART2 neural network. *Advanced Engineering Informatics, 16*, 229-240.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS quarterly, 36*, 1165-1188.

Chen, K., Luo, P., & Wang, H. (2017). An influence framework on product word-of-mouth (WoM) measurement. *Information & Management, 54*, 228-240.

Coleman, S., Göb, R., Manco, G., Pievatolo, A., Tort‐Martorell, X., & Reis, M. S. (2016). How can SMEs benefit from big data? Challenges and a path forward. Quality and Reliability Engineering International, 32(6), 2151-2164.

CoreNLP, S. (2016). a suite of core NLP tools. *URL http://nlp.stanford.edu/software/corenlp.shtml*.

Del Vecchio, P., Di Minin, A., Petruzzelli, A. M., Panniello, U., & Pirri, S. (2018). Big data for open innovation in SMEs and large corporations: Trends, opportunities, and challenges. Creativity and Innovation Management, 27(1), 6-22.

Demirkan, H., & Delen, D. (2013). Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. *Decision Support Systems, 55*, 412-421.

Dillon, T., Wu, C., & Chang, E. (2010). Cloud Computing: Issues and Challenges. In *2010 24th IEEE International Conference on Advanced Information Networking and Applications* (pp. 27-33).

Ghobadian, A., & Gallear, D. N. (1996). Total quality management in SMEs. *Omega, 24*, 83-106.

Goorha, S., & Ungar, L. (2010). Discovery of significant emerging trends. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 57-64): ACM.

Guo, D., Zhao, Y., Zhang, L., Wen, X., & Yin, C. (2019). Conformity feedback in an online review helpfulness evaluation task leads to less negative feedback-related negativity amplitudes and more positive P300 amplitudes. Journal of Neuroscience, Psychology, and Economics, 12(2), 73.

Hu, N., Liu, L., & Zhang, J. J. (2008). Do online reviews affect product sales? The role of reviewer characteristics and temporal effects. *Information Technology and management, 9*, 201-214.

Jun, S.-P., Park, D.-H., & Yeom, J. (2014). The possibility of using search traffic information to explore consumer product attitudes and forecast consumer preference. *Technological Forecasting and Social Change, 86*, 237-253.

Krueger, R., & Casey, M. (2009). Focus Groups, a Practical Guide for Applied Research 4 edition SAGE publications. *Inc, Thousand Oaks, California*.

Lewis, P., Thornhill, A., & Saunders, M. (2007). *Research methods for business students*: Pearson Education UK.

Li, J., & Zhan, L. (2011). Online persuasion: How the written word drives WOM: Evidence from consumer-generated product reviews. *Journal of Advertising Research, 51*, 239-257.

Lim, G. G., Kang, J. Y., Lee, J. K., & Lee, D. C. (2011). Rule-based personalized comparison shopping including delivery cost. *Electronic Commerce Research and Applications, 10*, 637-649.

Lim, S., & Tucker, C. S. (2016). A bayesian sampling method for product feature extraction from large-scale textual data. *Journal of Mechanical Design, 138*, 061403.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies, 5*, 1-167.

Liu, Y., Jin, J., Ji, P., Harding, J. A., & Fung, R. Y. K. (2013). Identifying helpful online reviews: A product designer's perspective. *Computer-Aided Design, 45*, 180-194.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55-60).

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity.

Maxwell, J. A. (2012). *Qualitative Research Design: An Interactive Approach: An Interactive Approach*: Sage.

Mourtzis, D., Vlachou, E., & Milas, N. (2016). Industrial Big Data as a Result of IoT Adoption in Manufacturing. *Procedia CIRP, 55*, 290-295.

Papacharalampopoulos, A., Stavridis, J., Stavropoulos, P., & Chryssolouris, G. (2016). Cloud-based Control of Thermal Based Manufacturing Processes. *Procedia CIRP, 55*, 254-259.

Philip Chen, C. L., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences, 275*, 314-347.

Russom, P. (2009). Next generation data warehouse platforms. *TDWI Best Practices Report: fourth quarter*.

Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. In *2013 International Conference on Collaboration Technologies and Systems (CTS)* (pp. 42-47).

Schmidt, R., & Möhring, M. (2013). Strategic Alignment of Cloud-Based Architectures for Big Data. In *2013 17th IEEE International Enterprise Distributed Object Computing Conference Workshops* (pp. 136-143).

Si, J., Li, Q., Qian, T., & Deng, X. (2014). Users' interest grouping from online reviews based on topic frequency and order. *World Wide Web, 17*, 1321-1342.

Soroka, A., Liu, Y., Han, L., & Haleem, M. S. (2017). Big data driven customer insights for SMEs in redistributed manufacturing. *Procedia CIRP, 63*, 692-697.

Stake, R. E. (1995). *The art of case study research*: Sage.

Sultan, N. A. (2011). Reaching for the "cloud": How SMEs can manage. *International Journal of Information Management, 31*, 272-278.

Tang, M., Jin, J., Liu, Y., Li, C., & Zhang, W. (2019). Integrating Topic, Sentiment, and Syntax for Modeling Online Reviews: A Topic Model Approach. *Journal of Computing and Information Science in Engineering, 19*, 011001.

Tucker, C., & Kim, H. (2011). Predicting emerging product design trend by mining publicly available customer review data. In *DS 68-6: Proceedings of the 18th International Conference on Engineering Design (ICED 11), Impacting Society through Engineering Design, Vol. 6: Design Information and Knowledge, Lyngby/Copenhagen, Denmark, 15.-19.08. 2011.*

Wan, Y., Menon, S., & Ramaprasad, A. (2007). A classification of product comparison agents. *Communications of the ACM, 50*, 65-71.

Wang, G., Xie, S., Liu, B., & Philip, S. Y. (2011). Review graph based online store review spammer detection. In *2011 IEEE 11th International Conference on Data Mining* (pp. 1242-1247): IEEE.

Ward, M., & Rhodes, C. (2014). Small businesses and the UK economy. *Report, House*.

Xu, X. (2012). From cloud computing to cloud manufacturing. *Robotics and Computer-Integrated Manufacturing, 28*, 75-86.

Yang, B., Liu, Y., Liang, Y., & Tang, M. (2019). Exploiting user experience from online customer reviews for product design. *International Journal of Information Management, 46*, 173-186.

Yin, R. K. (2003). Case study research: design and methods, Applied social research methods series. *Thousand Oaks, CA: Sage Publications, Inc. Afacan, Y., & Erbug, C.(2009). An interdisciplinary heuristic evaluation method for universal building design. Journal of Applied Ergonomics, 40*, 731-744.

Zhang, Z., Guo, C., & Goes, P. (2013). Product comparison networks for competitive analysis of online word-of-mouth. *ACM Transactions on Management Information Systems (TMIS), 3*, 20.