Journal of Services Mar

Text Mining Analysis Roadmap (TMAR) for Service Research

Journal:	*Journal of Services Marketing*
Manuscript ID	JSM-02-2019-0074.R2
Manuscript Type:	Article
Keywords:	Big Data, cluster analysis, Classification, qualitative research, artificial intelligence, content analysis

Text Mining Analysis Roadmap (TMAR) for Service Research

Abstract

Purpose: The purpose of this paper is to offer a step-by-step Text Mining Analysis Roadmap (TMAR) for service researchers. We provide guidance on how to choose between alternative tools, using illustrative examples from a range of different business contexts.

Design/methodology/approach: We provide a six staged Text Mining Analysis Roadmap on how to use text mining methods in practice. At each stage we: (1) provide a guiding question; (2) articulate the aim; (3) identify a range of methods; and (4) demonstrate how machine learning and linguistic techniques can be used in practice with illustrative examples drawn from business, from an array of data types, services and contexts.

Findings: At each of the six stages, we demonstrate useful insights that result from the text mining techniques to provide an in-depth understanding of the phenomenon and actionable insights for research and practice.

Originality/Value: There is little research to guide scholars and practitioners on how to gain insights from the extensive "big data" that arises from the different data sources. In a first, this paper addresses this important gap highlighting the advantages of using text mining to gain useful insights for theory testing and practice in different service contexts.

Key words: Text mining, service research, machine learning, natural language processing, qualitative research, Artificial intelligence

Paper Type: Research Paper

INTRODUCTION

Service researchers are faced with an unprecedened volume of textual data generated from a range of sources and formats such as research publications, news items, industrial reports, online chatter, surveys, interviews, blogs, scripts and notes. It is expected that the number and complexity of these qualitative data documents will only increase in the future. The International Data Group (IDG) predicts that by 2025 there will be 163 Zettabytes of data in the world, with around 80 percent of business-relevant information originating from unstructured forms, primarily text (Techrepublic 2017). Key sources of this growth are from firm-generated content including, for example, annual reports news items, and from customers' comments on websites and consumer-generated content that appear in social networking sites.

Methods of conducting and analyzing qualitative research are also changing. Data has become more readily available and tools such as text mining are well suited to handle large quantities of unstructured data and extract knowledge from these disparate primary and secondary data sources in short periods of time (Hartmann et al. 2016; Humphreys and Jen-Hui Wang 2017; McColl-Kennedy et al. 2019; Rust and Huang 2014; Villarroel Ordenes et al. 2014; and Zaki and Neely 2019). Text mining uses a set of natural language processing (NLP) and machine learning techniques to process textual documents, derive patterns within a structured format, and provide evaluation and interpretation of the output to gain insights that matter (Feldman 2006). It involves information retrieval, lexical analysis to study word frequency distributions, information extraction and machine learning techniques including visualization and predictive classification analytics (Schmunk et al. 2014). Yet despite its importance, these techniques, while well established in Information Systems and Computer Science literatures, are less well known in Service Research, and hence many service researchers do not know how to apply the techniques. Further, there is little research to guide service scholars and practitioners on how to determine which methods are most appropriate in different contexts.

Hence, the contribution of this article is three-fold; to: (1) provide a Text Mining Analysis Roadmap (TMAR) on how to use text mining methods in practice; (2) demonstrate the usefulness of text mining techniques to analyze qualitative data at the different research stages; (3) illustrate how service researchers can generate insights that result from the text mining techniques to provide a more in-depth understanding and also actionable insights for service marketing practice. At each of the six stages of our roadmap we: (1) provide a guiding question; (2) articulate the aim of the stage; (3) identify a range of methods; and (4) demonstrate how the machine learning and linguistic techniques can be used in practice with illustrative examples drawn from business, from a range of data types, services and contexts.

Importantly, we provide guidance on how to choose between the alternative methods providing illustrative examples from a range of different business contexts. This should enable service researchers and practitioners to generate insights providing in-depth understanding and actionable insights for practice.

Literature Review

In this first section we provide a review of Service Research studies that have used text mining, which are summarized in Table 1. As shown in Table 1, prior studies have focused on five topic areas: (1) literature reviews; (2) analyzing customer behavior; (3) branding and market structure; (4) online customer reviews; and (5) pricing. To illustrate we provide one example from each of these five areas.

Insert Table 1 about here

Anton and Breidbach (2018), Amado et al. (2017) and Mahr et al. (2019) used text mining techniques to automatically review and analyze literature from services, big data and marketing research. For example, Anton and Breidbach (2018) employed the Latent Dirichlet

Allocation (LDA) technique to extract topics from the textual publications (service innovation and service design) and tested the results using intercoder reliability (98%). They then folowed this up by assessing the labels and classifications with an external panel of researchers. The advantage of this approach is that it is able to review and synthesize large and heterogeneous bodies of work (such as journal publications, conference proceedings, web articles, books, book chapters and web articles). Future service studies could incorporate LDA and other text mining techniques to generate insights into other large scale data such as customer complaints, social media postings, predict churn rates (McColl-Kennedy et al. 2019; Rust and Huang 2014; Zaki 2019).

Regarding analyzing customer behavior, Villarroel Ordenes et al. (2014) and McColl-Kennedy et al. (2019) use text analytics to analyze customer feedback on key aspects of the customer experience. To illustrate, McColl-Kennedy et al. (2019) developed a customer experience (CX) conceptual framework that incorporates customer touchpoints, value creation elements, emotions and cognitive responses, and then applied the framework, using advanced text mining techniques, to two years of customer feedback in a complex B2B heavy asset service setting. They used a random sample of 100 comments from the dataset in the training stage (ranging in size from 246 to 255 characters) to provide rich text for data understanding and pattern development. They employed two coders who independently classified each comment following their conceptual framework. The text mining model uses part of speech (POS), and they developed macros and linguistic patterns rules applied to the conceptual framework. The resulting patterns were then mapped to the root causes, which enable the firm to identify opportunities to improve CX. They used chi-square automatic interaction detection (CHAID) classification technique to predict whether customers are satisfied. In short, their technique enables a firm to identify critical touchpoints from the *customer's* perspective – including potentially new touchpoints that had been previously unknown; understand what matters to the customer at each touchpoint; map each touchpoint to its root cause, i.e., the specific firm action or strategy; and take specific actions to improve the touchpoint and the overall customer experience. Further tests are needed to apply the model using different data sources, for example social media and other contexts such as B2C. Also, the LDA technique could be explored to automatically generate root cause topics in similar contexts.

Arvidsson and Caliandro (2016) and Tirunillai and Tellis (2014) used text mining to study branding and market structure. To illustrate, Tirunillai and Tellis (2014) extended the LDA technique to understand dimensions of product quality to gain insight into brand positioning. Using longitudinal data on product reviews across firms and markets, their study extracts specific latent dimensions of quality, and the valence, labels, validity, importance, dynamics, and heterogeneity of those dimensions. They used multiple methods (human raters and consumer reports) to validate the generated dimensions. However, the study is limited to product reviews and should be extended in future studies to include other forms of data such as news reports, financial documents, social media, online forum of products and other textual documents that marketing scholars often use. LDA is sensitive to values of hyperparameter and could influence the number of dimensions extracted.

Lee and Bradlow (2011), Netzer et al. (2012), Xiang et al. (2015) and Villarroel Ordenes et al. (2018), employed used text mining to study online customer reviews. To illustrate, Lee and Bradlow (2011) provide evidence of the power of text mining to identify and anlyze online product reviews that are easily repeatable for both physical products and services and that require minimal managerial intervention. After pre-processing the reviews, they used K-means clustering technique to extract different product attributes. However, they acknowledge that more complex topic-clustering, such as LDA and Latent semantic analysis could be considered.

Finally, regarding pricing Liu et al. (2018) developed a deep learning text-mining technique to extract quality and price content dimension from more than 500,000 reviews spanning nearly 600 product categories. After providing a topology and consumer review-reading behaviors, the authors quantify the causal impact of content information of read reviews on sales. They show that aesthetics and price content in the reviews significantly affect conversion across almost all product categories. Review content information has a higher impact on sales when the average rating is higher and the variance of ratings is lower. Consumers depend more on review content when the market is more competitive, immature, or when brand information is not easily accessible. This research could be extended by using text mining techniques to study the effect of reviews on consumer search behaviors.

In sum, this nascent stream of research demonstrates that there is growing interest in utilizing text mining models to analyze the qualitative data in service research. Yet, to date there is little to guide researchers and practitioners on how to use all the aforementioned text mining techniques on textual data generated from disparate data sources such as research publications, news, industrial reports, online chatter, or user-generated content, surveys, interviews, scripts, notes constitutes an excellent emerging source for service researchers. This is where our paper contributes.

Text Mining Analysis Roadmap (TMAR)

Our main goal is to provide a six staged roadmap for researchers to develop a text mining model (Figure 1). We adapted the Cross Industry Standard Process for Data Mining (Chapman et al. 2000; McColl-Kennedy et al. 2019; Villarroel Ordenes et al. 2014; Zaki and Neely 2019). The first stage – *Background Study* - involves generating common themes from the relevant publications. The second stage is the *Pre study: Business Understanding*, which comprises extracting information from secondary data (e.g. company reports, news, websites etc.) and primary data (e.g. observation notes or interview scripts, if available). The third stage, *Data Understanding,* involves the process of preparing the textual data for the modeling phase. The fourth stage, *Data Modeling*, includes the building and testing of the text mining model developed from the data understanding and processing data phase. Next, the *Data Validation* stage checks the reliability and accuracy of the text mining model and the final stage is the "*Insights Gained*" which enables interesting insights from the textual data to be gleaned.

Insert Figure 1 about here

At each step we: (1) articulate the aim; (2) provide a guiding question; (3) identify a range of techniques; and (4) demonstrate relative usefulness through illustrative examples in different contexts offering guidance on selecting from alternative techniques. We use the KNIME Analytics Platform for illustration purposes. This platform is an open source data science and machine learning platform that contains a text processing plugin to process textual and natural language data (Tursia and Silipo 2018). The platform has been used by service scholars Villarroel Ordenes et al. (2018). However, there are other text mining software programs and libraries (e.g. SAS, IBM SPSS Modeler, Python NLTK, R, as well as others) that can be employed. It is important to recognize that applying text mining techniques is an iterative process that requires tweaking parameters until reaching the best fit (Feldman 2006). Table 2 summarizes the text mining techniques used in this article and alternatives. We discuss the application of these techniques in the following sections, and provide illustrative examples from a range of different datasets and business contexts.

Insert Table 2 about here

Stage 1 Background Study

Aim. The aim of this stage is to generate themes from the literature review. The *guiding question* here is ***How can you generate key themes from your literature review?*** *Range of techniques*. Topic Modelling is one of the popular techniques in text mining which can automatically identify topics present in a text object to derive hidden patterns exhibited by a text corpus. Topics can be defined as a repeating pattern of co-occurring terms in a corpus (Humphrey and Wang 2017). Topic Models are useful for clustering, organizing large blocks of textual data and information retrieval from unstructured text into themes. They can be used to organize large datasets of emails, customer reviews, and user social media profiles. Topic Modelling is different from rule-based text mining approaches that use regular expressions or dictionary based keyword searching techniques. It is an unsupervised approach used for finding and observing the bag of words (BoW) (called "topics") in large clusters of texts. (Humphrey and Wang 2017).

There are many approaches to obtain topics or themes from a text such as Term Frequency (TF) and Inverse Document Frequency (IDF) and clustering techniques. The Latent Dirichlet Allocation (LDA) is the most popular topic modeling technique (Blei, Ng, and Jordan 2003) as LDA deals with a widely known problem, called data dimensionality (Pestov, 2013). In this section, we discuss how researchers can use it to automatically extract themes and topics from the literature. LDA assumes documents are produced from a mixture of topics. Those topics then generate words based on their probability distribution. LDA is a matrix factorization technique where a collection of publications can be represented as a document-term matrix. The following matrix shows a corpus of n documents (publications) D1, D2, D3 … Dn and vocabulary size of M words W1,W2 .. Wn. The value of i,j cell gives the frequency count of word Wj in Document Di.

An Illustrative Example. In this section we provide an illustrative example of the use of the LDA technique to generate key themes from 40 publications from service and customer experience research. These publications have to be pre-processed before any modelling can be undertaken. First, the text documents have to be parsed. Text mining software programs have a parser function that can read any text document of a certain format (e.g. PDF files). Second, a list of documents can then be used as input into the pre-processing step, where terms can be filtered and manipulated in order to remove terms that do not contain content, such as stop words, numbers, punctuation marks and convert it to lowercase in order to eliminate ambiguity with uppercase words. Finally, researchers can define the ideal number of topics from the literature. It can be difficult to estimate the ideal number of topics. Techniques such as the Elbow method can be used to cluster the data to find the optimal number of clusters before using LDA. Another approach is to use Gibbs sampling to estimate the LDA model (i.e., the posterior distribution of the topics) (Griffith and Steyvers 2004). In our case, we identified 10 topics (Topic_0 to Topic_9) to classify the overall topic landscape of currently published service and customer experience research.

To illustrate, we extracted the top 10 terms (using word weight matrix) associated with each topic as well as the titles, abstracts, key words and content of all articles on the respective topic, and classified all articles with a meaningful topic as shown in Figure 2. However, while LDA identifies the individual topics in a text corpus, the associated terms, and the topic loadings of all documents, it does not generate a label to describe each topic (Blei 2012). The labeling of LDA outputs still need human input (Antons and Breidbach 2018; Bao and Datta 2014). Following Antons and Breidbach's (2018) lead, we recommend that at least two researchers independently label the topics based on their own judgments. Figure 2 shows that the 40 articles sampled generated 10 research themes: customer experience in different contexts (Topic_0, Topic_4 and Topic_8), such as retail, healthcare and hospitality services; customer experience management and measurement (Topic_1, Topic_2 and Topic_6); service systems model and service design innovation (Topic_3, Topic_4 and Topic_7). Finally, there are topics on the impact of social media on the next generation of consumer market research (Topic_5 and Topic_9). This machine learning approach should help service researchers to address gaps in service research knowledge and reveal hidden structures, and describe new development themes. Furthermore, it provides a holistic picture and assessment of the research area for future research (Antons and Breidbach 2018; Gallouj and Savona 2009; Mahr et al. 2019).

Insert Figure 2 about here

Stage 2 Pre-study Business Understanding

Aim. The overall aim of any pre-study is to gain business understanding. This can be undertaken through identifying key secondary data sources (e.g. websites, news, company reports, etc.) which could help in the study prior to collecting the primary data sources (e.g. observations, interviews, focus group, experiments, workshops, surveys, etc.). The *guiding question* at this stage is ***How can you analyze secondary data and primary data?***

We recommend that researchers familiarize themselves with the service and the context. The aim here is to show how to extract meaningful themes and knowledge from secondary data sources. Often researchers need to source and extract data from numerous websites to create datasets, usually a website offers application programming interfaces (APIs) which enable structured data to be "fetched". There are a range of techniques incuding: (1) web scrappping; (2) feature selection; (3) feature representation; and (4) clustering techniques.

Web scraping. There are many web scraping functions available in most of the text mining software programs. For example, in KNIME there is a package called Palladian (http://www.palladian.ws/). The package reads a given online Hypertext Markup Language (HTML) website and extracts all information in further processable Extensible Markup Language (XML) format. Similarly, IBM SPSS Modeler has a web feed node which can retrieve information from the web. Python scripting language has a web scraping library, called "beautifulsoup " that extracts specific information such as structured data from the HTML/XML content. R software for statistical computing has a "rvest" package, to scrape data from html web pages which returns an XML document containg all the information about the web page.

Feature selection. The decision regarding generating features from the text (bottom-up/unsupervised or top-down/supervised) is crucial because if there is incorrect input, the tools will populate a meaningless output. There are many feature selection techniques to help with this task. The most common technique is called bag of words (BoW) which in simple terms means breaking the text up into words and considering each of them as a feature (Nassirtoussi et al. 2014). Thus, after parsing and processing the press releases corpus, the documents are transformed into a BoW which is filtered again. Only terms that occur at least in 1% of the documents are used as features and not be filtered out.

Feature representation. After the minimum number of features is determined, each feature needs to be represented by a numeric value so that it can be processed by machine learning algorithms called “feature-representation”. This assigned numeric value acts like a score or weight. We calculated the most widely used frequency measures in text mining, that is, the term frequency (tf) and inverse document frequency (idf) to specify the relevancy of a term. In the case of the term frequency tf *(t,d),* the simplest choice is to use the raw count of a term in a document, i.e., the number of times that term t occurs in document d. If we denote the raw count by $f_{t,d}$ then the simplest tf scheme is

$$tf(t,d) = f_{t,d}$$

The Inverse Document Frequency (idf) describes the importance of the term calculated by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient. *N*, the Total number of documents in corpus $N = |D|$, $\{d \in D : t \in d\}$: the number of documents where the term *t* appears.

$$\text{idf(t,D)} = \log \frac{N}{|\{d \ \in D : t \ \in d\}|}$$

Then we can calculate *tf-idf* to measure a term weight that is intended to reflect how important a word is to a document in a corpus.

$$tfidf\ (t,d,D) = tf(t,d).\ idf\ (t,D)$$

Other techniques can be used such as *n-grams* (Butler and Kešelj 2009; Hagenau et al. 2013), *Chi-square keyword extractor* and *Keygraph keyword extractor* (Beliga et al. 2015). An n-gram is a contiguous sequence of n items which are usually words from a given sequence of text. After the frequencies have been computed, the terms can be filtered, related to the frequency values that are required for the analysis. It is possible to specify minimum and maximum values or a particular number k, so that only the k terms with the highest frequency are retained. Beside extracting keywords based on their *tf-idf* value, other techniques can be applied as well, such as the "Chi-square keyword extractor" using a chi-square measure to score the relevance (Beliga et al. 2015) or the "Keygraph keyword extractor" (Beliga et al. 2015) using a graph representation of documents to find keywords to describe the document.

Clustering. After the pre-processing is completed and text is transformed into a number of features with a numeric representation, cluster algorithms can be tested (e.g. hierarchical clustering, K-means, K-medoids and others). But first researchers need to use distance functions to determine the relationship between data points. Corresponding with the intended outcome of our cluster analysis the k-medoids clustering algorithm was selected. The k-medoids algorithm groups *n* objects into *k* clusters by minimizing the sum of dissimilarity between each object, *p*, and its corresponding representative object, o_i (medoid), for all objects in cluster *Ci* (Han et al. 2011):

$$\min\ E = \sum_{i=1}^{k} \sum_{p\ \in\ \mathrm{C}i} dist(p,o_i)$$

Selecting the number of clusters is a trade-off between having a reasonably large number of clusters to reflect the specific differences in the dataset, and having significantly fewer clusters than data points (Han et al. 2011). Cluster analysis requires researchers to interpret the clusters in order to provide insights that are meaningful.

An llustrative Example. We collected press releases about customer experience from a sample of multinational companies listed in The Fortune Global 500. The data was collected from the business information and research search engine Factiva (www.factiva.com). Factiva collects content from both licensed as well as free sources online and provides access to over 32,000 sources including newspapers, journals, press releases, magazines, photos, and television and radio transcripts (Dow Jones, 2017).

The search query used the terms 'customer experience' + 'selected Fortune 500 company'. Results were filtered for data in June 2017. All documents returned based on the search query were then downloaded in a text file format (RTF). In total, the search query returned 10,770 documents for a total of 230 companies. In this section, we demonstrate how to apply document clustering techniques such as k-Means, k-Medoids or hierarchical clustering of a sample of a total of 179 documents on 230 Fortune Global 500 companies (financial services, fast-moving consumer goods (FCMG), energy, aerospace and defense, healthcare, logistics, media, motor vehicles and parts, oil and gas, retail, IT Technology, telecommunication and travel) are collected. Many studies successfully improve text mining results by adding a Named Entity Recognition (NER) model to targeted companies (e.g. Vu et al. 2012). We applied OpenNLP name entity tag embeded in KNIME software to automatically extract the organizations' names. Researchers can filter documents based on the different verticals to extract insights.

We used the customer experience frameworks suggested by (Bolton et al. 2018; Lemon and Verhof 2016) to shed light on the four clusters – (1) Behavioral; (2) Emotional;

(3) Cognitive; and (4) Relational. The first cluster consists of documents from 30 firms on designing *behavioral* experiences. For example, DHL Parcel is expanding its customer service to include a new voice-activated information service. DHL customers can query Amazon's digital smart speaker "Alexa" for information on their parcel's current whereabouts. The second cluster consists of 47 firms that design *emotional* experiences (Bolton et al. 2018). An example of a peak experience is Coca-Cola Company together with 12 artists from around the world and Delta creating an art gallery in the sky, transforming the tray tables on one of the airline's 767 aircraft into original works of art. Each piece of art celebrates themes of optimism, travel, refreshment and happiness. The third cluster is *cognitive* and consists of 26 firms. For example, BMW is the first manufacturer to develop Augmented Reality (AR) to enhance the dealership experience. BMW's customers can download the app to their smartphone and then hold the device in front of them to visualize the vehicle and explore the interior and exterior of the car, interact with its features (such as the car's lights or trunk), and customize its color and wheel rims with the tap of the screen.

The fourth cluster is *relational* consisting of 29 firms. For example, the Ghana Association for the Deaf has given rave reviews about Vodafone's SuperCare" initiative – which delivers a specialized customer service for the speech and hearing-impaired individuals. Vodafone designed the first customer experience center for speech and hearing impaired, providing a unique service that takes its social responsibility a few notches higher. There has been no special package innovatively designed service until now for people with speech and hearing impairment.

We recommend that researchers validate the text mining analysis by collecting primary data. For example, researchers could visit key functional units and shadow employees from each unit. Detailed field notes, photographic records, and memos could be undertaken. Following established protocol for qualitative data analysis (Glaser and Strauss

1967; Spiggle 1994), text mining models could also be used to generate categories from these data. Second**,** in addition to the observations and shadowing, interviews could be conducted with key service actors (e.g. employees, customers, service providers, etc.). Semi-structured interviews provide an effective means of obtaining rich insights into the phenomenon of interest along with items that can be analyzed using quantitative methods (McCracken, 1988). Combining these research protocols should enable interesting results and insights.

Stage 3 Data Understanding

Aim. The aim of this stage is to prepare and commence processing the textual data. The *guiding question* here is ***How can you prepare and process textual data sources?*** There are many different textual data sources that are generated from marketing systems and platforms such as surveys, CRM, social media and many others (Zaki and Neely 2019). We demonstrate below how researchers can prepare and process these textual data using a top-down approach employing theoretically-guided methods (Humphrey and Wang 2017).

A *range of techniques* is available to help expand theory and make discoveries by allowing coders to independently classify each comment into discrete units of information (Singh, Hillmer and Wang 2011). The labeling (manual coding) process should be undertaken by at least two coders (e.g. researchers or using crowdsourcing with platforms such as Mechanical Turk) and the inter-rater reliability measured (Miles and Huberman 1994; Rust and Cooil 1994). When there is disagreement, the two trained coders should discuss their coding with a third person, who acts as a judge (e.g. one of the research authors) until a decision is reached (Brady, Voorhees and Brusco 2012; Fastoso and Whitelock 2010). Scott's Pi coefficient should be calculated for each dimension using the following formula:

$$\text{Scott's Pi } = \frac{Po - Pc}{1 - Pc}$$

Scott's Pi formula calculates the coefficient of percent agreement (Po) and the percent of expected agreement (Pc). A coefficient above 0.80 is considered an acceptable level of reliability, whereas a coefficient above 0.90 indicates a high level of reliability (Riffe, Lacy and Fico 1998). Some studies label the full dataset, while others (e.g. McColl-Kennedy et al. 2019; Villarroel Ordenes et al. 2014; Zaki and Neely 2019) suggest taking a random sample of 100 comments from the dataset for the training stage and the full dataset used for the testing stage.

The top-down approach could be used to: (1) generate descriptive insights from the data; (2) develop a linguistic text mining model that can automatically extract concepts based on manual coding and built-in analyzers and dictionaries; and (3) offer guidance for the text mining model validation which compares between each generated (predicted) field and its target field (labeled).

An Illustrative Example. We downloaded twitter data from *www.Kaggle.com* which is a platform for sharing open data sources and running machine learning competitions in which companies and researchers compete to produce the best machine learning and AI models for predicting and describing the data (Kaggle 2019). We chose the airline industry (Business to Consumer) because it is a challenging service business environment with great interest to academics (Gilbert and Wong, 2003; Ostrowski et al. 1993; Pakdil and Aydin 2007). Further, buying a ticket is not a frequent purchase, and if the flight is cancelled, emotions are heightened.

Consequently, social media provides customers with a means to express disappointment publicly and quickly. Premium airlines such as British Airways, who have created their reputation based on customer service, are at higher risk of social media meltdown when an accident occurs (Lauchlan, 2017; McEleny, 2017; Shearman, 2017). For example, British Airways experienced an IT system failure in 2017, leaving 75,000 passengers stranded in Gatwick and Heathrow airports. The crisis worsened due to lack of timely communication with customers on social media, and failure of its CEO to make a public comment on the accident. The incident occurred a month after United Airlines underwent a social media catastrophe, following the publishing of a video showing a passenger being forcefully removed from an overbooked flight. Also, United Airlines received considerable negative comments on social media, after it held back two women wearing leggings from boarding a flight. This resulted in the *#leggingsgate* appearing on Twitter. Moreover, Delta had to handle customer complains on social media, after defending airline overbooking, and hundreds of flight cancellations due to IT glitches (Lauchlan, 2017; McEleny, 2017; Shearman, 2017). Clearly, social media data is a rich data source to gather and analyze real-time comments from customers and obtain information about the service. Our dataset consists of 15,591 comments on February 2015 from Twitter of American, United, Southwest, British Airways, JetBlue, US Airways and Virgin America airlines.

Service researchers can use either a top-down approach or a bottom-up such as supervised and unsupervised machine learning techniques (Humphrey and Jen-Hui Wang 2017). In this section, we highlight the top-down approach, where researchers can follow a manual coding process using software such as Microsoft Excel or NVivo software through reading each verbatim comment and manually categorizing the comment through applying a conceptual framework. In our case (Table 2), the comments were annotated according to four dimensions generated from CX literature on customer evaluation as either (positive or negative or neutral) (McColl-Kennedy et al. 2012), customer journey (pre-purchase, purchase, post-purchase) (Lemon and Verhoef 2016), touchpoints (brand, customer, partner, external) (Lemon and Verhoef 2016; McColl-Kennedy et al. 2019), and root cause categorization (McColl-Kennedy et al. 2019).

Insert Table 3 about here

Researchers can generate descriptive statistics. This analysis demonstrates that most comments in the pre-purchase phase are neutral (51%), which may be the result of customers' information about the service. Another interesting pattern can be seen in the purchase and post-purchase stages as they are constructed mainly as negative comments, 69% and 71% respectively, showing great potential for improvement. Most of the touchpoints in the pre-purchase stage belonged to customers (49%), as at this point they were looking for information about the service. Moreover, most of the touchpoints in the purchase and post-purchase stages were brand-owned, 68% and 60% respectively, showing that firms can have a high impact on the creation of CX throughout the customer journey. Partner and external-owned touchpoints constituted a small percent of comments. However, they should be investigated and monitored as they can contribute to word-of-mouth. The last part of the analysis was based on categories of touchpoints. They found that the most significant elements of airline services as viewed by customers, are punctuality, price, customer service (online, phone system and in-person), luggage transportation, and loyalty programs. Not surprisingly, the main pain points relate to delays, cancelled flights, lost luggage, in-flight services, and bookings (e.g. Bejou et al. 1996; Gursoy et al. 2005; Pakdil and Aydin, 2007).

This example shows that insights can be generated from the data understanding stage. First, we can demonstrate that social media can be used to build positive brand awareness. Interestingly, JetBlue and Southwest use Twitter to communicate about new destinations. Moreover, the team shares updates related to on-time flights every time a flight arrived at its destination sooner than expected. Second, we can analyze the impact of a firm's intervention. As discussed in literature, fast intervention by the customer service team in response to complaints means that concerned customers can receive help immediately. This may strengthen their loyalty, and build a solid and long-lasting relationship. For example, a Virgin America passenger tweeted the airline after his luggage has been stolen saying that he was stuck at the airport, wearing the same clothes for a couple of days. The social media team responded quickly and offered help to the passenger. The passenger tweeted back about his experience, spreading positive word-of-mouth to other customers. Finally, examining touch points from social media can assist in predicting and decreasing the churn rate. Our analysis showed that customers threatened to leave the firm after experiencing a service failure. Specifically, as shown, 18% of customers decided to leave before accomplishing the purchase (pre-purchase), largely due to issues with the website, application, purchase, and lack of information or assistance.

Stage 4 Data Modeling

Aim. In this section, we demonstrate how researchers can use a bottom-up classification approach to analyze survey verbatim comments of a B2B heavy asset service that offers both physical goods and services. The *guiding question* at this stage is ***How can you develop a text mining model?***

A *range of techniques* are available including (1) parsing documents; (2) enrichment and tagging process; and (3) classification machine learning models. Regarding *parsing documents*, text documents have to be parsed (analyzed into logical syntactic components). Text Mining software has a parser function that can read text documents of certain formats, such as DML, SDML, PubMed (XML format), PDF, Microsoft files, CSV, flat files and other formats. In the *enrichment and tagging process,* parts of speech (POS) tagging can be used to assign part of speech tags (e.g. Stanford tagger, Abner tagger). This is not limited to English language. Most Text Mining software programs have taggers that can work with other languages, such as, German, French, Spanish, Chinese. Researchers can create their own domain-specific dictionary. Third, in the pre-processing step, terms can be filtered and manipulated in order to remove terms that do not contain content such as stop words, numbers, punctuation marks or prepositions words, such as “to”, or “at”. Terms are also filtered to remove endings based on declination or conjugation by applying stemming. Regular expression is another powerful text pre-processing function that can be utilized for linguistic purposes. For example, we can remove all digits in the document or exclude comments that have characters, such as, $, £, %, #, @. We can use the bag of words (BoW) technique where a piece of text (sentence or a document) is represented as a bag of words, disregarding grammar and even word order and the frequency or occurrence of each word is used as a feature (numerical vector) for training a machine learning model.

Regarding *classification of machine learning models,* after the enrichment and tagging pre-processing are completed and text is transformed into a number of features with a numeric representation and machine learning algorithms can be engaged. Further, n-gram features could be used in addition to single words to take into account negations, such as "not good" or "not bad". For the classification we can use techniques such as Decision Rules or Trees, Naïve Bayes, Support Vector Machine (SVM), Regression Algorithms, Neural Networks, Combinatory Algorithms or Multi-algorithm experiments (Nassirtoussi et al. 2014).

For the decision rule classifier, a set of decision rules can suggest a pattern of words found in the documents. Decision trees are special decision rules that are organized into a tree structure which divides the document space into non-overlapping regions at its leaves, and predictions are measured at each leaf (Weiss et al., 2010). Naïve Bayes is rooted in applying Bayes' theorem and is "called naïve because it is based on the naïve assumption of complete independence between text features" (Nassirtoussi et al. 2014). This algorithm examines the feature space to calculate the "posterior probability that a document belongs to different classes and assigns it to the class with the highest posterior probability" (Tan and Zhang 2008). Unlike Naïve Bayes, SVM is example-based (Gupta et al. 2013). The idea behind SVMs is locating a "decision surface" that is used to separate the training instances, for example, into positive and negative examples, the support vectors being a few of these instances which act as the “only effective elements in the training set” (Tan and Zhang 2008). SVM is particularly well-suited to classification tasks which involve a large number of features (Ravi and Ravi 2015). Multilayer perceptron (MLP) consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. MLP utilizes a technique called backpropagation for training. A recent development in neural network is the creation of a library called Keras which is an open source neural network written in Python and it is integrated into many text mining software programs like KNIME Analytics platform. It is capable of running on top of TensorFlow and the Microsoft Cognitive Toolkit which is designed to enable fast experimentation with deep neural networks focusing on being user-friendly, modular, and extensible (https://keras.io).

*An Illustrative example***.** As an example, consider that a customer survey is administered on a monthly basis (including structured and unstructured data) to evaluate customer satisfaction. The survey includes 12 questions and the final question is an open-ended question, “Do you have any other comments or suggestions on how (NAME) could improve this service”. This allows the company to gather real-time comments from customers and obtain information about the service not elsewhere captured. Our data consists of 2650 responses over a 12 month period. We used BoW and POS tagging techniques to perform sentiment analysis (opinion mining) to extract subjective information from a piece of text (positive or negative) (Nassirtoussi et al. 2014). For example, as shown in Figure 3 in the following quote, “The technician failed to complete the servicing in time” would carry a negative sentiment. Further, the extent of the negativity is assessed. For example, “The service was slow and the technician was rude. I would expect a lot better, the experience was unsatisfactory” would be assigned a greater negative sentiment than the previous comment. For sentiment analysis, the tags need to be converted to strings for analysis. These strings are then converted to a number format to allow a quantitative summation to be used to assign sentiments.

Insert Figure 3 about here

This stage could be followed by a binning process which can automatically assign bins to the comments based on their sentiment. Four bins were used, ranging from positive to very positive (1 to 2) or negative to very negative (-1 to -2). In our analysis, sentiment labels, such as positive (1) or negative (-1) are assigned automatically to each comment. Furthermore, each comment was weighted to measure if the customer's evaluation is skewed toward positive or negative sentiment. For example, when the comment has more positive terms than the negative terms, this could be given a very positive label (2). Using the same logic when customers complained more, the weight of negative terms would be assigned a very negative label (-2).

As in all supervised mining algorithms, we need a target variable (e.g. sentiment label). With classification algorithms, a sample of the data as a training set (70%) and a test set (30%) should be used. In the test dataset, these patterns and correlations were used in the classification phase to predict the sentiment of the unused data (30%). Researchers could experiment with different algorithms or a combination of multiple algorithms until reaching the best results. In our case, the decision tree technique outperformed other algorithms and produced accuracy of 94%.

Stage 5 Data Validation

Aim. According to an extensive review of text mining studies by Nassirtoussi et al. (2014), most of the studies report the "accuracy, recall or precision and sometimes the F-measure" for the purpose of evaluating the developed models. The *guiding question* at this stage is ***How can you validate a text mining model?***

Range of techniques. Accuracy is the most commonly reported figure, and in most cases it is between 50 – 80%. The review by Nassirtoussi et al. (2014) concludes that accuracies of above 55% are accepted as "report-worthy", but also point out that the majority of the reviewed studies do not comment on whether the data used was "imbalanced" or not. This is important as an imbalanced dataset can strongly affect the accuracy measure and not reporting the data structure therefore raises doubts about a model's performance. The below formulas are used to calculate recall (R), precision (P), the F-measure (F1) and accuracy which is the most common measure (Nassirtoussi et al. 2014). True positives are labelled as *tp*, false positives as *fp*, true negatives as *tn*, and false negatives as *fn*.

$$\text{Recall (R)} = \frac{tp}{tp + fn}$$

$$\text{Precision (P)} = \frac{tp}{tp + fp}$$

$$\text{F Measure (F1)} \; = \frac{2 \times P \times R}{P + R}$$

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

Cross validation is used to assess the variance in the model's classification performance. A 10-fold cross validation means that the datasets were split into 10 equally sized subsets. Cross validation is a powerful general technique to estimate a model's performance and has been shown to be a better estimator of it than testing (Seni and Elder 2010).

As a further check, researchers could employ a linguistic graduate assistant to manually check and validate the text mining model output. Following McColl-Kennedy et al. (2019), Singh, Hillmer and Wang (2011) and Villarroel Ordenes et al. (2014), the linguistic graduate assistant should be given detailed instructions on the coding scheme. The coder's role is to compare the text mining output and check each comment manually identifying whether the text mining model mapped the comments/words correctly to a relevant topic/class.

*An Illustrative example***.** For example, to verify the accuracy of the sentiment analysis in the previous section, the sentiment scores produced by the text mining should be compared to the overall satisfaction for a given survey response.

Stage 6 Insights Gained

Aim. In this section, we demonstrate how service researchers can use visualization techniques to generate deep insights from the textual data. This section could be considered the managerial implications part. The *guiding question* at this stage is ***How can you generate insights from textual data?*** *Range of techniques*. Text mining software and tools have built-in techniques and extensions for visualization libraries. Alternatively, researchers can export the generated model outcome into a tabular format (e.g. CSV, Excel, etc.) and then import it into another visualization software program (e.g. Tableau).

*An Illustrative example***.** Figure 4 shows a visualization of a live service mapping that is a geographical representation of the customer service responses from B2B survey data. This can be built by adding geographic positions of services from the survey dataset. We combined quantitive and qualitative measures to test whether there is correlation between topics generated from the verbatim comments and the quantitative measures, such as overall satisfaction and referral scores. For example, we plotted the overall customer satisfaction scores versus the customers' root causes from the verbatim comments. We can check the average of overall customer satisfaction for certain categories of root cause to identify which geographical areas have "complaints" and "compliments". The color scale shows which services are performing well or unsatisfactory from the referral scores and customers' comments. The annual service ratings are another useful means to analyze the historical performance versus the root cause categories. For example, an annual loyalty variations plot can be overlaid onto a box-whisker plot to identify anomalies across the longitudinal dataset. These graphs can be filtered by year, month, week, even daily. This could identify months which appear to be particularly problematic, or identify a root cause which is more common at a certain time of year than at other times.

Insert Figure 4 about here

Conclusion

In this article, we make at least two important contributions. First, we respond to calls for work to demonstrate the usefulness of employing using text mining techniques to the ever growing amount of qualitative data in service research (e.g. Anton and Breidbach 2018; Humphrey and Wang 2017; McColl-Kennedy et al. 2019; Villarroel Ordenes et al. 2018; and Zaki 2019), showing how to use text mining in practice across a range of contexts. Second, we provide a six staged Text Mining Analysis Roadmap (TMAR) to guide researchers, including offering practical guidance on how to choose between the alternatives through illustrative examples from a range of different business contexts.

In the first stage, we demonstrated how researchers can generate key themes from literature reviews using topic model techniques such as LDA. In the second stage, we discussed different approaches to guide researchers in the pre-study and business understanding stages, including: (1) how to fetch structured data from many websites (using web scraping functions) and press releases; (2) how to generate features from the text (bottom-up/unsupervised or top-down/supervised) such as ‘‘bag of words’’ and ‘‘feature-representation” such as term frequency (tf), inverse document frequency (idf), “Chi-square keyword extractor” and Keygraph keyword extractor; and (3) how to use different cluster algorithms (e.g. hierarchical clustering, K-means, K-medoids and others) to cluster secondary data documents. In the third stage, we showed how researchers can prepare and process textual data (e.g. airline twitter data) using a top-down approach (theoretically-guided methods) to: (1) generate descriptive insights from the data; (2) develop a linguistic text mining model that could automatically extract concepts based on manual coding and built-in analyzers and dictionaries; and (3) offer guidance for the text mining model validation which compares between each generated (predicted) field and its target field (labeled). In the fourth stage, we demonstrated how researchers can use a bottom-top (classification) approach to analyze survey verbatim comments illustrating this with data from a B2B heavy asset firm and evaluate customer satisfaction using classification machine learning (e.g. Decision Rules or Trees, Naïve Bayes, Support Vector Machine (SVM), Regression Algorithms, Neural Networks, and Combinatory Algorithms or Multi-algorithm experiments). In the fifth stage, we explained how to use accuracy, recall (R) or precision (P), the F-measure and human coders to evaluate and valu evaluating and valididate the developed text mining models. Finally, we demonstrated how researchers can generate insights by visualization (e.g. a live service mapping and root cause analyzes). Going forward we encourage service researchers to not only use the techniques discussed above but also to investigate new techniques as they become available from Computer Science and related fields.

REFERENCES

Amado, A., Cortez, P., Rita, P., and Moro, S. (2017), "Trends on big data in marketing: A text mining and topic modeling based literature analysis", *European Research on Management and Business Economics*, Vol. 24, No. 1, pp. 1–7.

Antons, D. and Breidbach, C. F. (2018), "Big data, big insights? Advancing service innovation and design with machine learning", *Journal of Service Research*, Vol. 21, No.1, pp.17–39.

Arvidsson, A. and Caliandro, A. (2016), "Brand public", *Journal of Consumer Research*, Vol. 42, No.5, pp. 727-748.

Bao, Y.and Anindya, D.(2014), "Simultaneously discovering and quantifying risk types from Textual Risk Disclosures," *Management Science*, Vol. 60, No. 6, pp. 1371-1391.

Beliga, S., Mestrovic, A., Martincic-Ipsic, S. (2015), "An overview of graph-based keyword extraction methods and approaches", *Journal of Information and Organizational Sciences*, Vol. 39, No.1, pp. 1-20.

Bejou, D., Edvardsson, B., and Rakowski, J. P. (1996) "A critical incident approach to examining the effects of service failures on customer relationships: The case of Swedish and U.S. airlines", *Journal of Travel Research*, Vol. 35, No.1, pp. 35–40.

Blei, D. M. (2012), "Probabilistic topic models", *Communications of the ACM*, Vol. 55, No. 4, pp. 77-84.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003), "Latent Dirichlet Allocation" *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022.

Bolton, R. N., McColl-Kennedy, J. R., Cheung, L., Gallan, A., Orsingher, C., Witell, L., and Zaki, M. (2018), "Customer experience challenges: bringing together digital, physical and social realms"*, Journal of Service Management,* Vol. 29, No. 5, pp. 776-808.

Brady, M. K., Clay M. V. and Brusco, M. J. (2012), "Service sweethearting: Its antecedents and customer consequences", *Journal of Marketing*, Vol. 76, No. 2, pp. 81-98.

Büschken, J., and Allenby, G. M. (2016), "Sentence-based text analysis for customer reviews", *Marketing Science*, Vol. 35, No. 6, pp. 953–975.

Chapman, P., Clinton, J., Kerber, R. and Khabaza, T. (2000), "CRISP-DM 1.0 Step-by-step data mining guide", available at: ftp://ftp.software.ibmcom/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf (accessed 30 January 2019).

Dow Jones (2017), "Global news database: Licensed content", available at: https://www.dowjones.com/products/factiva/ (accessed 20 December 2018).

Fastoso, F. and Whitelock, J. (2010), "Regionalization vs. globalization in advertising research: insights from five decades of academic study", *Journal of International Management*, Vol. 16, No. 1, pp. 32-42.

Feldman, R. and Sanger, J. (2006), *The Text Mining Handbook*. Cambridge University Press, New York, NY.

Gallouj, F. and Savona, M. (2009), "Innovation in services: A review of the debate and research agenda," *Journal of Evolutionary Economics*, Vol. 19, No. 2, pp. 149-172.

Glaser, B.G. and Strauss, A. L. (1967), *The Discovery of Grounded Theory.* Aldine, Chicago.

Gilbert, D. and Wong, R. K. C. (2003), "Passenger expectations and airline services: A Hong Kong based study", *Tourism Management*, Vol. 24, No.5, pp. 519–532.

Griffiths, T. L. and Mark, S. (2004), "Finding scientific topics", *Proceedings of the National Academy of Sciences*, pp. 5228-5235.

Gupta, N., Gilbert, M. and di Fabbrizio, G. (2013), "Emotion detection in email customer care", *Computational Intelligence*, Vol. 29, No. 3, pp. 489–505.

Gursoy, D., Chen, M. H. and Kim, H. J. (2005), "The US airlines relative positioning based on attributes of service quality", *Tourism Management*, Vol. 26, No. 1, pp. 57–67.

Han, J., Kamber, M. and Pei, J. (2011), *Data Mining Concepts and Techniques*, Elsevier Science, Burlington.

Hartmann, P. M., Zaki, M., Feldmann, N. and Neely, A. (2016), "Capturing value from big data - a taxonomy of data-driven business models used by start-up firms", *International Journal of Operations & Production Management*, Vol. 36, No.10, pp. 1382-140.

Humphreys, A. and Rebecca J. W. (2017), "Automated text analysis for consumer research," *Journal of Consumer Research*, Vol. 44 No. 6, pp. 1274-1306.

Kaggle (2019), available at: https://www.kaggle.com (accessed 28 January 2019).

Lauchlan, S. (2017) "United's social media nightmare leaves Delta CEO to defend airline overbooking", available at: https://diginomica.com/uniteds-social-media-nightmare-leaves-delta-ceo-defend-airline-overbooking/ (accessed 19 January 2019).

Lee, T.Y. and Bradlow, E. T. (2011), "Automated marketing research using online customer reviews", *Journal of Marketing Research*, Vol. 48, No. 5, pp. 881-894.

Lemon, K. N. and Verhoef, P. C. (2016), "Understanding customer experience throughout the customer journey", *Journal of Marketing*, Vol. 80, No. 6, pp. 69-96.

Liu, X., Lee, D. and Srinivasan, K. (2018), "Large-scale cross-category analysis of consumer review content on sales conversion leveraging deep learning", Marketing Science Institute Working Paper Series.

Mahr, D., Stead, S. and Odekerken-Schröder, G. (2019), "Making sense of customer service experiences: a text mining review", *Journal of Services Marketing*, Vol. 33, No.1, pp. 88-103.

McColl-Kennedy, J. R., Vargo, S., Dagger, T.S , Sweeney, J. C. and Kasteren, Y.V. (2012), "Health care customer value cocreation practice styles," *Journal of Service Research*, Vol. 15, No.4, pp. 370-389.

McColl-Kennedy, J. R., Zaki, M., Lemon, K. N., Urmetzer, F. and Neely, A. (2019), "Gaining customer experience insights that matter", *Journal of Service Research*, Vol. 21, No.1, pp. 8-26.

McCracken, G. (1988), *The Long Interview*. SAGE Qualitative Research Methods Series 13*,* Beverley Hills.

McEleny, C. (2017), "Airlines' engagement with customers on social media is now critical to success", available at: https://www.thedrum.com/opinion/2016/11/25/airlines-engagement-with-customers-social-media-now-critical-success (accessed 18 January 2019).

Miles, M. and Huberman, M. (1994), *Qualitative Data Analysis: An Expanded Sourcebook*. Sage, Thousand Oaks, CA.

Netzer, O., Ronen F., Jacob G. and Moshe F. (2012), "Mine your own business: Market-structure surveillance through text mining," *Marketing Science*, Vol. 31, No.3, pp. 521-43.

Nassirtoussi, K., A., Aghabozorgi, S., Ying Wah, T. and Ngo, D. C. L. (2014), "Text mining for market prediction: A systematic review", *Expert Systems with Applications*, Vol. 41, No.16, pp. 7653–7670.

Pakdil, F. and Aydin, Ö. (2007), "Expectations and perceptions in airline services: An analysis using weighted SERVQUAL scores", *Journal of Air Transport Management*, Vol. 13, No. 4, pp. 229–237.

Pestov, V. (2013), "Is the -NN classifier in high dimensions affected by the curse of dimensionality?" *Computers and Mathematics with Applications*, Vol. 65, No. 10, pp.1427–1437.

Ravi, K. and Ravi, V. (2015), "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications", *Knowledge-Based Systems*, Vol. 89, pp. 14–46.

Riffe, D., Lacy, S. and Fico, F. (1998), *Analyzing Media Messages: Using Quantitative Content Analysis in Research*. Erlbaum, New York, NY.

Rust, R. T. and Bruce, C. (1994), "Reliability measures for qualitative data: Theory and implications," *Journal of Marketing Research*, Vol. 31, No. 1, pp. 1-14.

Rust, R. T. and Ming-Hui, H. (2014), "The service revolution and the transformation of marketing science," *Marketing Science*, Vol. 33, No.2 , pp. 206–21.

Schmunk, S., Wolfram, H., Matthias F., and Maria L. (2014), “Sentiment analysis: Extracting decision relevant knowledge from UGC” in *Information and Communication Technologies in Tourism 2014*, pp. 253-265.

Seni, G. and Elder, J. (2010), *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions*, McMorgan and Claypool Publishers, Chicago.

Shearman, S. (2017), “Why airlines need a handle on social media to build their brand”, available at: https://www.ft.com/content/45459474-46dd-11e7-8d27-59b4dd6296b8 (acessed 10 November 2019).

Singh, S. N., Steve, H. and Ze, W. (2011), “Efficient methods for sampling responses from large-scale qualitative data,” *Marketing Science*, Vol. 30, No.3,pp. 532-549.

Spiggle, S. (1994), “Analysis and interpretation of qualitative data in consumer research,” *Journal of Consumer Research*, Vol. 21, No.4, pp. 491-503.

Tan, S. and Zhang, J. (2008). “An empirical study of sentiment analysis for Chinese documents”, *Expert Systems with Applications*, Vol. 34, No. 4, pp. 2622–2629.

Techrepublic (2017), “Unstructured data: A cheat sheet”, available at: https://www.techrepublic.com/article/unstructured-data-the-smart-persons-guide/ (accessed 9 August 2019).

Tirunillai, S. and Gerard, J. T. (2014), “Mining marketing meaning from online chatter: Strategic brand analysis of big data using Latent Dirichlet Allocation”, *Journal of Marketing Research*, Vol. 51, No. 4, pp. 463-479.

Vincenzo, T. and Rosaria, S. (2018), *From Words to Wisdom: An Introduction to Text Mining with Knime*, Zurich, Switzerland: KNIME Press.

Ostrowski, P. L., O’Brien, T. V. and Gordon, G. L. (1993), “Service quality and customer loyalty in the commercial airline industry”, *Journal of Travel Research*, Vol. 32, No. 2, pp. 16–24.

Villarroel Ordenes, F., Theodoulidis, B., Burton, J., Gruber, T. and Zaki, M. (2014), “Analyzing customer experience feedback using text mining: A linguistics-based approach,” *Journal of Service Research*, Vol. 17, No. 3, pp. 278-295.

Villaroel Ordenes, F., Grewal, D., Ludwig, S., Ruyter, K. D., Mahr, D., Wetzels, M. and Kopalle, P. (2019), “Cutting through content clutter: How speech and image acts drive consumer sharing of social media brand messages”, *Journal of Consumer Research*, Vol. 45, No. 5, pp. 988–1012.

Vu, T. T., Chang, S., Ha, Q. T. and Collier, N. (2012), *“An experiment in integrating sentiment features for tech stock prediction in twitter”* in *Proceedings of the workshop on information extraction and entity analytics on social media data* Mumbai, India, pp. 23–38.

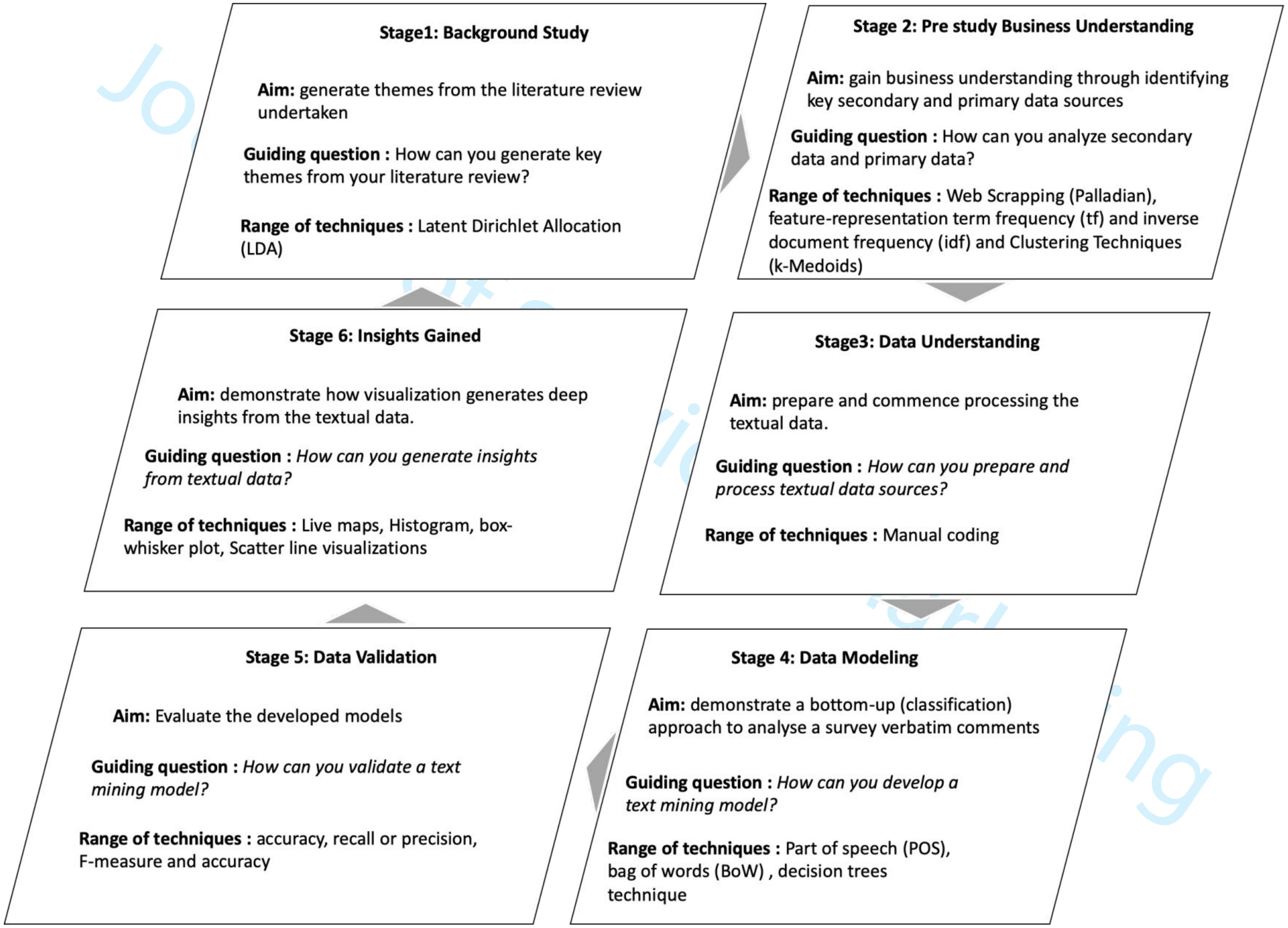
Weiss, S. M., Indurkhya, N. and Zhang, T. (2010). Fundamentals of Predictive Text Mining.

Xiang, Z., Zvi S. , John H. G. and Muzaffer U. (2015), “What can big data and text analytics tell us about hotel guest experience and satisfaction?”, *International Journal of Hospitality Management*, Vol. 44, pp. 120–130.

Zaki, M. and Neely, A. (2019), “*Customer experience analytics: Dynamic customer-centric model*”. In: Maglio, P.P., Kieliszewski, C.A., Spohrer, J.C., Lyons, K., Sawatani, Y. and Patrício, L. (2 eds.). *Handbook of Service Science*. Volume 2, Springer International Publishing AG.

Zaki, M. (2019), "Digital transformation: Harnessing digital technologies for the next generation of services," *Journal of Service Marketing*, Vol. 33, No. 4, pp. 429-435.

Figure 1 Text Mining Analysis Roadmap (TMAR)



Stage1: Background Study

Aim: generate themes from the literature review undertaken

Guiding question : How can you generate key themes from your literature review?

Range of techniques : Latent Dirichlet Allocation (LDA)

Stage 2: Pre study Business Understanding

Aim: gain business understanding through identifying key secondary and primary data sources

Guiding question : How can you analyze secondary data and primary data?

Range of techniques : Web Scrapping (Palladian), feature-representation term frequency (tf) and inverse document frequency (idf) and Clustering Techniques (k-Medoids)

Stage3: Data Understanding

Aim: prepare and commence processing the textual data.

Guiding question : *How can you prepare and process textual data sources?*

Range of techniques : Manual coding

Stage 4: Data Modeling

Aim: demonstrate a bottom-up (classification) approach to analyse a survey verbatim comments

Guiding question : *How can you develop a text mining model?*

Range of techniques : Part of speech (POS), bag of words (BoW) , decision trees technique

Stage 5: Data Validation

Aim: Evaluate the developed models

Guiding question : *How can you validate a text mining model?*

Range of techniques : accuracy, recall or precision, F-measure and accuracy

Stage 6: Insights Gained

Aim: demonstrate how visualization generates deep insights from the textual data.

Guiding question : *How can you generate insights from textual data?*

Range of techniques : Live maps, Histogram, box-whisker plot, Scatter line visualizations

Figure 2 Topic Modeling using Latent Dirichlet Allocation (LDA) Technique to Generate Themes

Customer experience and emotion in healthcare

Topics	Terms	Weight
topic_0	emot	331
topic_0	health	260
topic_0	activ	242
topic_0	custom	238
topic_0	experi	234
topic_0	care	203
topic_0	well-being	173
topic_0	research	155

Customer experience management in service firms

Topics	Terms	Weight
topic_1	custom	2111
topic_1	experi	1327
topic_1	market	797
topic_1	servic	716
topic_1	research	652
topic_1	manag	568
topic_1	cx	429
topic_1	firm	403

Customer experience measurement in retail(physical and digital)

Topics	Terms	Weight
topic_2	model	339
topic_2	experi	326
topic_2	onlin	299
topic_2	custom	266
topic_2	consum	207
topic_2	store	180
topic_2	retail	169
topic_2	measur	163

Service systems model

Topics	Terms	Weight
topic_3	model	426
topic_3	system	315
topic_3	studi	254
topic_3	tion	249
topic_3	user	226
topic_3	time	198
topic_3	compani	190
topic_3	comput	183
topic_3	peopl	177

Service design innovation in hospitality

Topics	Terms	Weight
topic_4	servic	1944
topic_4	innov	799
topic_4	research	781
topic_4	design	602
topic_4	topic	260
topic_4	fals	258
topic_4	true	195
topic_4	food	165

Social media and the new generations choice and commitment

Topics	Terms	Weight
topic_5	social	389
topic_5	vol	273
topic_5	choic	221
topic_5	media	204
topic_5	commit	164
topic_5	gen	146
topic_5	social	389

Service technology (digital and social): customer behavior and interaction with employees

Topics	Terms	Weight
topic_6	servic	579
topic_6	technologi	344
topic_6	custom	320
topic_6	employe	284
topic_6	behavior	253
topic_6	research	208
topic_6	consum	203
topic_6	interact	184
topic_6	social	183

Service value: processes, capabilities and resources

Topics	Terms	Weight
topic_7	servic	614
topic_7	custom	546
topic_7	provid	326
topic_7	valu	251
topic_7	process	204
topic_7	resourc	157
topic_7	capabl	151
topic_7	activ	136
topic_7	mine	122

Customer experience design in retail

Topics	Terms	Weight
topic_8	experi	842
topic_8	custom	620
topic_8	design	425
topic_8	retail	355
topic_8	product	344
topic_8	research	236
topic_8	aspect	225
topic_8	engag	219
topic_8	market	209

The effect of social media on consumer market research

Topics	Terms	Weight
topic_9	market	1135
topic_9	research	530
topic_9	onlin	481
topic_9	social	453
topic_9	consum	431
topic_9	custom	303
topic_9	media	285
topic_9	scienc	272
topic_9	effect	262

Figure 3 Sentiment Analysis: Transforming Qualitative Measures to Quantitative Measures

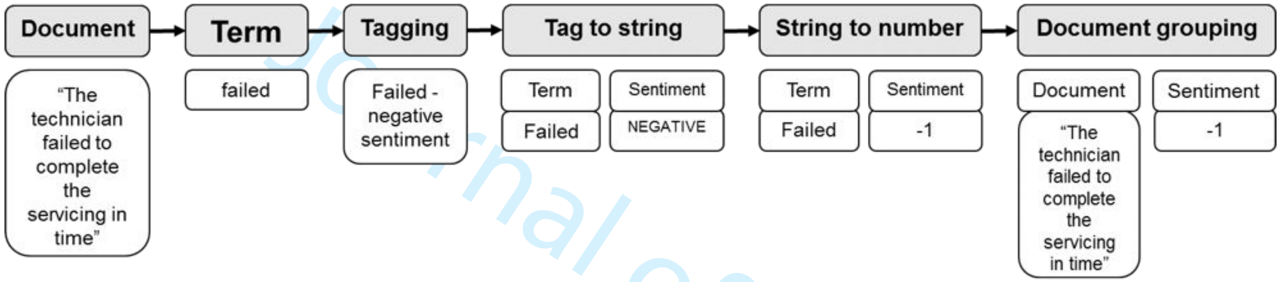
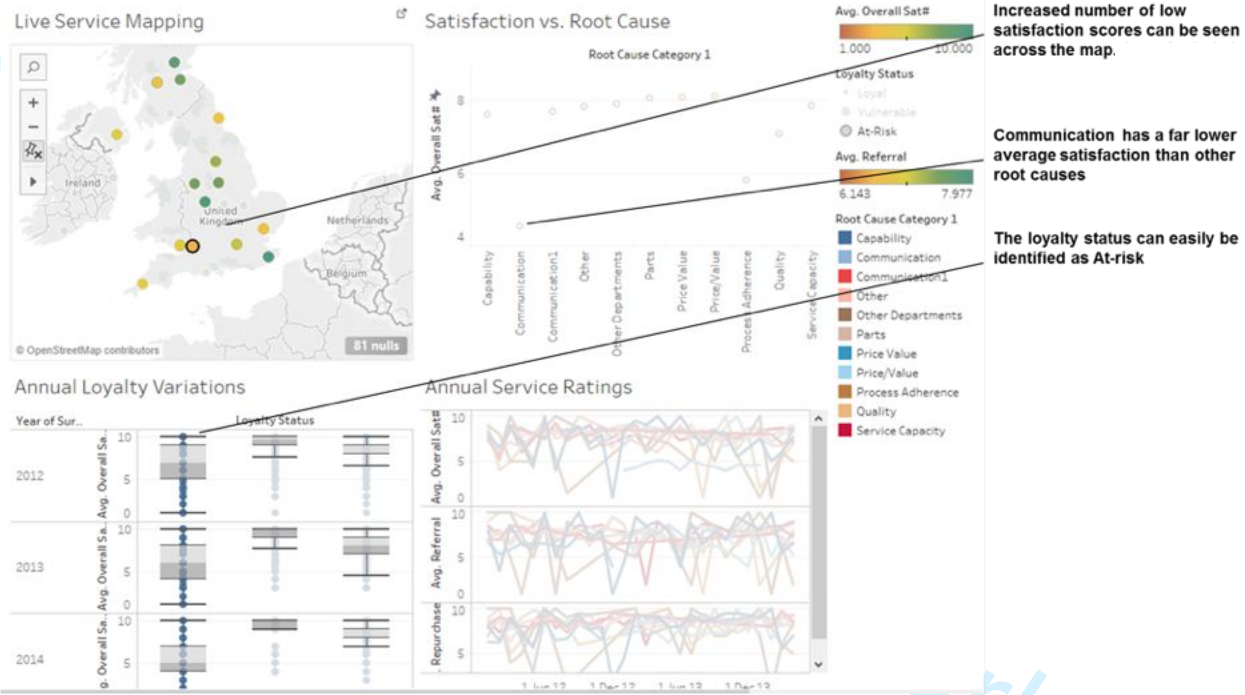

Document
"The technician failed to complete the servicing in time"
Term
failed
Tagging
Failed - negative sentiment
Tag to string
Term
Sentiment
Failed
NEGATIVE
String to number
Term
Sentiment
Failed
-1
Document grouping
Document
Sentiment
"The technician failed to complete the servicing in time"
-1

Figure 4 Customers' Evaluation Visualizations



Increased number of low satisfaction scores can be seen across the map.

Communication has a far lower average satisfaction than other root causes

The loyalty status can easily be identified as At-risk

Table 1 A Review of Text Mining Literature in Service Research Studies

Topics	Authors	Datasets	Text Mining Techniques and Tools	Key Findings
Literature review	Mahr et al. (2019)	Analyze articles from services and marketing journal (n=258 articles)	Leximancer Text Mining Software	• There is a shift from brands and products to value and interaction, with a stronger focus on emotional and sensorial dimensions in the physical servicescapes
	Anton and Breidbach (2018)	Analyze articles from service innovation and service design research (n = 641 articles)	Latent Dirichlet Allocation (LDA)	• Extract 69 distinct research topics • Portray the entire network of 69 topics using network analysis to visualize the research topics • Produce an extensive research agenda consisting of four research directions and 12 operationalizable guidelines to service research
	Amado et al. (2017)	Analyze articles from big data in marketing research (n = 1560 articles)	Latent Dirichlet Allocation (LDA)	• Generate a matrix structure with the frequency of terms in these documents • Analyze the three most significant terms based on distribution • Big data publications in Marketing literature are not clearly aligning with cutting edge big data techniques • Big data applications in Marketing is still in the early stages with most of the research originating from Europe, North America and Asia
Analyzing customer behavior	McColl-Kennedy et al. (2019)	Analyze customers' survey of Asset heavy B2B firm (n=3,116 comments)	Coding schema, dictionary, linguistic patterns and CHAID classification techniques	• Identify critical touchpoints from the customer's perspective – including potentially new touchpoints that had been previously unknown • Understand what specifically matters to the customer about each touchpoint • Map each touchpoint to its root cause, i.e., the specific firm action or strategy • Take specific actions to improve the touchpoint and the overall customer experience (CX)
	Villarroel Ordenes et al. (2014)	Analyze customers' feedback of car parking and transfer services (n= 1,092 comments)	Coding schema, dictionary and linguistic patterns	• The highest percentage of complaints and compliments pertained to the use of the facilities to park cars • The analysis classified various complaints about company resources, such as signage, space, staff, facilities, etc. • 47 subcategories have been generated to provide a clear picture of the parking and transfer service process • Differentiate between the activities performed by the customers and those performed by the company • Classify and map 678 concepts to these subcategories • Differentiate between ''personal context'' and ''situational context'' that can affect customers' experiences
Branding and market structure	Arvidsson and Caliandro (2016)	Analyze social media data (n= 8949 tweets about Louis Vuitton)	Coding schema, co-occurrence analysis, social network analysis	• Brand publics are social formations that are not based on interaction but on a continuous focus of interest and mediation • Brand publics might be part of a social media-based consumer culture where publicity rather than identity has become a core value • In brand publics consumers do not develop a collective identity around the focal brand; rather the brand is valuable as a medium that can offer publicity to a multitude of diverse situations of identity • Brand publics might be part of a social media–based consumer culture where publicity rather than identity has become a core value
	Tirunillai and Tellis (2014)	Analyze customers' reviews from five markets (n=350,000 reviews)	Latent Dirichlet Allocation (LDA)	• Identification of specific latent dimensions of quality, and the valence, labels, validity, importance, dynamics, and heterogeneity of those dimensions • The model analyzes the dynamics of experienced quality at a highly granular temporal level • Demonstrates the importance of extracted dimensions by the time-varying intensity of the conversions on each dimension • Demonstrates how brands compete on multidimensional space and varies over time in a great detail • The dimensions of quality can be a basis for determining consumer satisfaction, brand ranking, new product design and content design
Online customer reviews	Villarroel Ordenes et al. (2018)	Facebook posts (n= 12,374 posts) and Twitter (n= 29,413 tweets)	Regular expression, coding schema and a support vector machine (SVM) classifier	• Directive messages, or explicit calls to action, induce less consumer sharing than assertive (informational/factual) or expressive (emotional) messages • Messages with socioemotional intentions are more likely to be exchanged • Facebook's function as a more emotional social network. On Twitter, assertive or factual content is more readily exchanged, and messages aim to spread information • Assertive and expressive (cf. directive) messages that feature alliteration trigger greater consumer sharing on social media platforms • Posting the same message type (e.g., assertive followed by another assertive) reduces consumer engagement, whereas complementarity and varied cross-message compositions result in greater message sharing
	Xiang et al. (2015)	Analyze customers' reviews from 10,537 hotels from Expedia.com (n= 60,648 customer review)	Coding schema, dictionary and factor analysis based on co-occurrences of words describe hotel experience	• There is a strong association between aspects of the customer experience (hygiene and motivation factors) and satisfaction • A guest who stays with family members seems to be not interested in the service aspect of the hotel other than a spacious room and nearby attractions • Identification of the Family Friendliness factor, which shows that what the guest brings into the experience can be an important contributing factor to their satisfaction • Domain knowledge proved to be critical in guiding the data processing and analytical process before reaching the point where meaningful relationships emerged
	Netzer et al. (2012)	Analyze online user-generated content from Edmunds' car forums (n= 868,174 comments) and pharmaceutical drugs (n=671,102 comments)	Co-occurrences of product, supervised machine learning architectures such as CRFs with rule-based or dictionary-based text mining and semantic network analysis	• The analyzed car market structure from the forum is highly correlated with the market structure derived from traditional data • Cars that share similar characteristics and/or are similarly mentioned with respect to terms referring to the luxuriousness of the car, the value consumers receive from it, its appearance, and the emotional sentiment mentioned about the car are more likely to be mentioned together in a message. • Text mining can assess the competitive market structure through consumers' perceptions of the products' attributes (car problems or ADRs)

	Lee and Bradlow (2011)	Analyze reviews from Epinions.com (2004-2008)	LDA and Latent semantic analysis	• The analysis can compare the brand's position with a focus on the salient product attributes and reveals underlying segments to evaluate the success of a campaign's objectives and consumer perception
Pricing	Liu et al. (2018)	Analyze 600 product categories from Amazon (n=500,000 reviews)	Deep learning	• Aesthetics and price content in the reviews significantly affect conversion across almost all product categories • Review content information has a higher impact on sales when the average rating is higher and the variance of ratings is lower • Consumers rely more on review content when the market is more competitive, immature, or when brand information is not easily accessible

Table 2 Summary of Text Mining Techniques - Examples Across the TMAR Stages

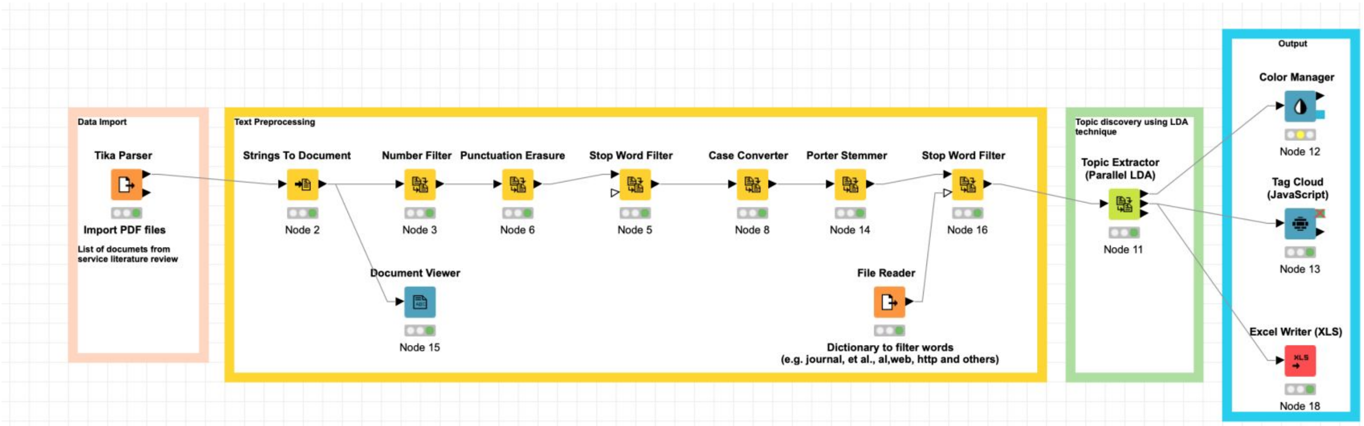
TMAR Stages	Datasets	Illustrative techniques used in this article	Alternative techniques
Stage 1 **Background Study**	40 publications from service and customer experience research	Latent Dirichlet Allocation (LDA)	Term Frequency (TF) and Inverse Document Frequency (IDF) and clustering techniques
Stage 2 **Pre-study Business Understanding**	Press releases from Factiva (www.factiva.com). about customer experience. A sample of 10,770 documents for a total of 230 multinational companies listed in The Fortune Global 500.	Web scraping techniques: KNIME- Palladian package Feature selection techniques: Bag-of-words Feature representation techniques: TF-IDF Clustering techniques: K-medoids	Web scraping techniques: For example, Beautifulsoup in Python, rvest in R Software and Web feed in IBM SPSS Modeler. Feature selection techniques: n-grams Feature representation techniques: Chi-square keyword extractor and Keygraph keyword extractor Clustering techniques: Hierarchical clustering and K-means
Stage 3 **Data Understanding**	Airline twitter data (15,591 tweets) for one month, February 2015	Top-down approach: Manual Coding and inter-rater reliability measure	Bottom-up approach: N-grams, BoW, TF-IDF, POS tagging, regular expression, Latent Dirichlet Allocation (LDA), Sentence-Constrained (SC) LDA and Sticky SC-LDA and Clustering techniques
Stage 4 **Data Modeling**	Survey verbatim comments of a B2B heavy asset service (2650 responses over 12-month period	Decision Trees	Naïve Bayes, Support Vector Machine (SVM), Regression Algorithms, Neural Networks, Combinatory Algorithms or Multi-algorithm experiments and deep neural networks
Stage 5 **Data Validation**	Survey verbatim comments of a B2B heavy asset service (2650 responses over 12-month period)	Accuracy and compare the qualitative measures with quantitative measures (e.g. overall satisfaction)	Recall (R), precision (P), the F-measure (F1), 10-fold cross validation and human assistant
Stage 6 **Insights Gained**	Survey verbatim comments of a B2B heavy asset services (2650 responses over 12-month period)	Live maps, Histogram, Box-whisker plot, Scatter line visualizations	Other built-in techniques and extensions for visualization libraries in text mining software

Table 3 Labeling (Manual Coding) the Twitter Dataset

Airline	Twitter Comments	Sentiment	Customer Journey Stage	Touchpoints	Root Cause
United	@united Gate agent hooked me up with alternate flights. If you have a way to PREVENT the constant issues, that would rock	Negative	Postpurchase	Brand	Gate agent
JetBlue	@JetBlue how can your entire system go down? No IT? Really?	Negative	Prepurchase	Customer	Website
British Airways	@British_Airways Our flights to SF cancelled Sat. Rebooked with Virgin. In-bound BA flights need to remain. Confirm plse	Neutral	Purchase	Partner	Cancelled flight
US Airways	@USAirways still can't get a real person on the phone to book a flight. Ready to just go with #jetblue since they care. #usairwayssucks	Negative	Purchase	Partner	Call Center Booking
US Airways	@USAirways Oh certainly. And now I have two $275 pending transactions on my bank account. Really happy that I was charged double	Negative	Purchase	External	Charge
American	@AmericanAir @justynmoro I totally agree. You get the automatic phone attendent that goes NO WHERE and hangs up. Lousy service!	Negative	Postpurchase	External	Phone system
British Airways	More great stuff from @helenbevan understands #Leadership #HR tnx 2 @IanHesky 4 RT #leaders #listen #hear #learn no… https://t.co/L3MCSqs6tJ	Positive	Prepurchase	External	Advertisement

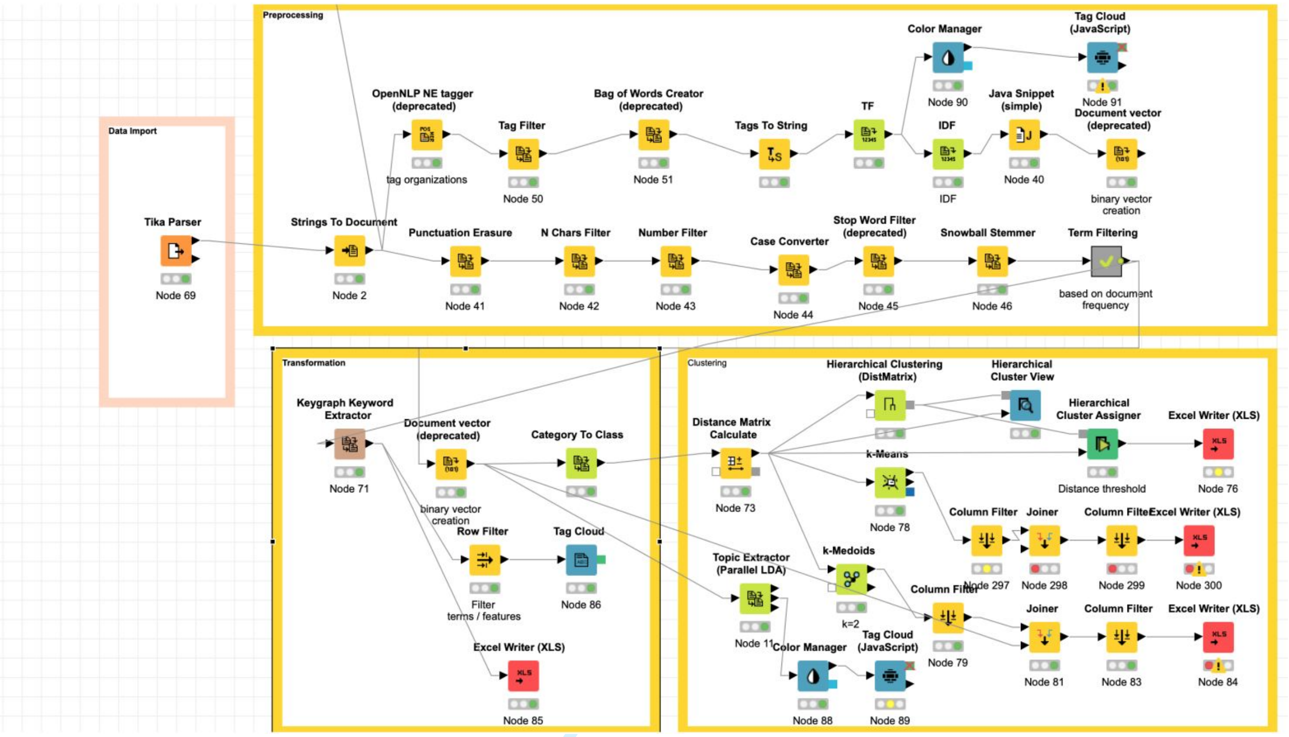
Appendix 1

Topic modelling workstream to extract themes from literature:



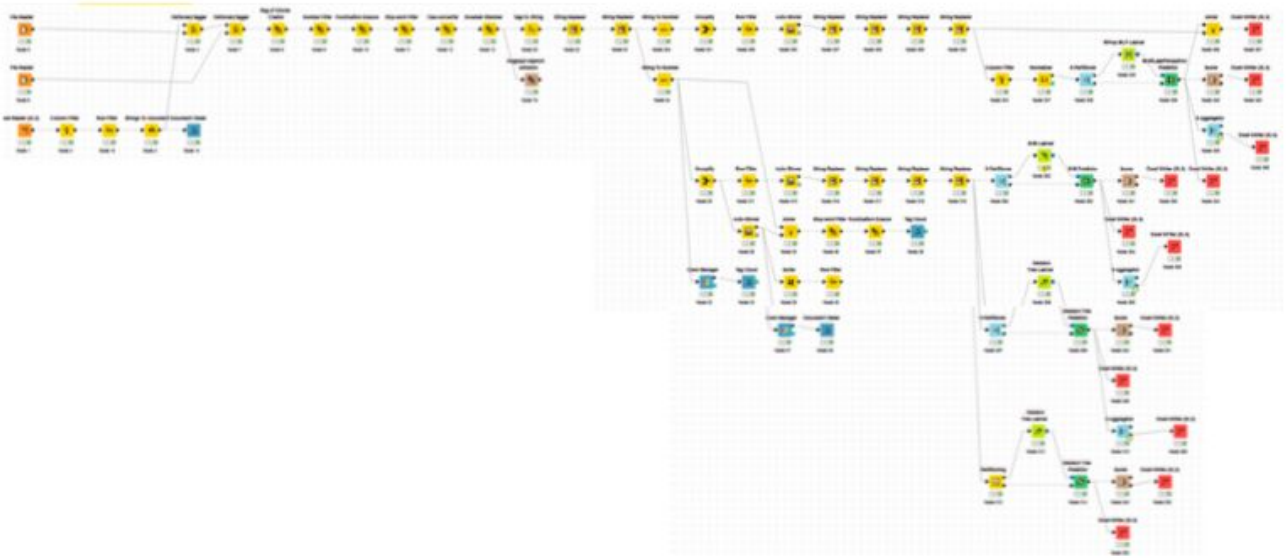
Appendix 2

Clustering Documents from Factiva press releases



Appendix 3

Classification Models- An example of Sentiment Analysis



Appendix 4

The most common Text Mining libraries in Python

Library	Description
NLTK	a platform for building Python programs to work with textual data. It provides easy-to-use interfaces to lexical resources such as WordNet.It also has text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning.
Gensim	a library for topic modelling, document indexing and similarity retrieval with large corpora.
spaCy	a commercial open source software. It has a pipeline for fast, state-of-the-art natural language processing.
Polyglot	a natural language pipeline that supports massive multilingual applications. It supports tokenization for 165 languages, Language detection for 196 languages, Named Entity Recognition for 40 languages, Part of Speech Tagging for 16 languages, Sentiment Analysis for 136 languages, Word Embeddings for 137 languages, Morphological analysis for 135 languages, transelation for 69 languages
Scikit-learn	a machine learning library which has a huge number of algorithims (e.g. regression, SVM, decision trees, etc.)
Kersa (TensorFlow)	The high-level neural network library (deep learning library) that developed by Google. It is a high-level neural network library that helps in designing the network architectures while avoiding the low-level details.