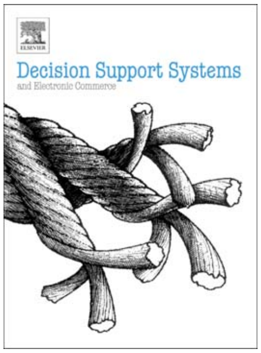

# A Computational Model for Mining Consumer Perceptions in Social Media

Demitrios E. Pournarakis, Dionysios N. Sotiropoulos, George M. Giaglis

PII: S0167-9236(16)30167-1
DOI: doi: 10.1016/j.dss.2016.09.018
Reference: DECSUP 12775

To appear in: *Decision Support Systems*

Received date: 8 December 2014
Revised date: 16 September 2016
Accepted date: 24 September 2016

# A Computational Model for Mining Consumer Perceptions in Social Media

Demitrios E. Pournarakis[a], Dionysios N. Sotiropoulos[a], George M. Giaglis[a]

[a] *Department of Management Science & Technology, Athens University of Economics & Business, Greece. Address: Elpidos 13, 10434*

## Abstract

The proliferation of Big Data & Analytics in recent years has compelled marketing practitioners to search for new methods when faced with assessing brand performance during brand equity appraisal. One of the challenges of current practices is that these methods rely heavily on traditional data collection and analysis methods such as questionnaires, and face to face or telephone interviews, which have a significant time lag. In this paper we introduce a computational model that combines topic and sentiment classification to elicit influential subjects from consumer perceptions in social media. Our model devises a novel Genetic Algorithm to improve clustering of tweets in semantically coherent groups, which act as an essential prerequisite when searching for prevailing topics and sentiment in big pools of data. To illustrate the validity of our model, we apply it to the Uber transportation network, from data collected through Twitter for the period between January and April 2015. The results obtained present consumer perceptions and produce insights for two fundamental brand equity dimensions: brand awareness and brand meaning. Simultaneously, they improve clustering results, in comparison to the k-means approach.

*Keywords:* Social Media, Big Data, Consumer Perceptions

## 1. Introduction

Reaching into consumers' minds and extracting knowledge about their experience during consumer - brand interaction has been one of the dominating areas of research in marketing. This concept of ability to experience, identify or understand customer attitudes towards brands derives from the notion of empathy [1] introduced by psychologists in the 19th century. In their attempts to conceptualize this notion, marketing researchers focus attention on the construct of brand equity which refers to the added value consumers place on a product or service [2] and has been most comprehensively defined by Aaker, Keller and Berry [3, 4, 5] as part of their studies during the 90s.

Marketers constantly seek ways to justify the impact of online and offline marketing activities in terms of brand performance in the marketing mix. With financial measures providing only partial indicators of brand performance, marketers turn towards assessment of intangible market based assets such as brand equity [6]. Several researchers [7, 8, 9, 10] have conceptualized brand equity similarly to Aaker and Keller and have presented fundamental dimensions that govern its nature.

One of the challenges of current practices that try to assess these dimensions is that they rely heavily on traditional data collection and analysis and thus have a significant time lag. However, recent years have witnessed a fundamental shift in how consumers choose to convey their experience during their interaction with a brand. The rise of social media platforms has empowered consumers worldwide to influence and shape perceptions of brands, leaving no choice for organizations than to adapt, invest in and monitor these channels. With mobile growth driving the change towards a connected consumer who with the power of a smartphone now has access to information previously only available through conventional means, marketing and technology are, more than ever, drawing towards a merge. Information is now digitized and stored - in the form of news, blogs, forum posts and social networks - making it increasingly important not to neglect such means [11].

Motivation for this paper arose from the need to introduce a new method for eliciting influential factors that govern brand equity assessment, by mining and analyzing consumer perceptions from online social network data. To do so we apply a design science research approach. We describe the design and evaluation of a computational model that lays out a proposed method on how consumer perceptions could be detected, starting from a marketing perspective. We follow that with the technical steps necessary to construct the assessment metrics, and conclude with how the results could be interpreted and utilized in brand equity assessment exercises.

This study aims to contribute to the literature in the following ways. Firstly we contend that traditional methods of data gathering provide a static and sometimes skewed indication of customer perceptions. On the other hand, assessment through Social Media Analytics (SMA) provides a dynamic and near real time measure of what customers are expressing at the moment of brand interaction. Secondly, this paper investigates how marketing and information systems can be intertwined (both as scientific disciplines and within the organizational context) to compute, assess and interpret customer attitudes towards a brand.

From a practitioner's perspective, we provide managers with an actionable method that can be utilized and applied in daily operations. In this sense, insights originating from social media channels may act as complementary measures to existing practices, justifying specific marketing strategies or brand performance results.

To illustrate the validity of our model, we apply it to the Uber transportation network. We collected and analyzed a set of over 280.000 tweets, during a three-month period *(January - April 2015)*, by utilizing the Streaming API of Twitter. The data collection process was focused on gathering tweets that were explicitly referring to Uber by performing hash-tag and mention filtering on the terms "#uber" and "@uber".

The remainder of this paper is organized as follows. In section 2, we review current approaches of brand equity assessment and list computational methods that are currently used to solve complex problems of data analysis from online sources. Our proposed model is introduced in section 3 and the results of the case study are presented in section 4. In section 5 we discuss the results and conclude by proposing future directions in section 6.

## 2. Theoretical Background

The brand as a concept has been the subject of much research within the marketing community, especially with regard to the construct's equity [3, 4, 5]. Attempts to provide brand equity assessment methods, based on the consumer perspective, have seen light in marketing literature during the past two decades. Prevailing methods focus on the multi-dimensionality of the construct, using confirmatory factor analysis with structural equations modeling for evaluating specific dimensions [12, 13, 14, 15]. Table 1 summarizes studies that are widely used during brand equity assessment along with the dimensions of brand equity upon which each study draws.

Although these measurement approaches have been and continue to be

| **Authors** | **Brand Equity Dimensions** |
|---|---|
| Aaker (1992) | Brand Loyalty, Brand Awareness, Perceived Quality, Brand Associations, Other Brand Assets |
| Berry (2000), Keller (1993) | Brand Awareness, Brand Meaning |
| C. J. Cobb-Walgren et al. (1995), Pappu et al. (2005), Washburn & Plank (2002), Yoo & Donthu (2001), Veloutsou and Christodoulides (2013), Tong & Hawely (2009) | Brand Awareness, Brand Associations, Perceived Quality, Brand Loyalty |
| Christodoulides et al. (2006) | Emotional Connection, Online Experience, Responsive Service nature, Trust, Fulfillment |
| French et al. (2013) | Brand Associations |
| Lassar et al (1995) | Attachment, Performance, Social Image, Trust, Value |

Table 1: Brand Equity Dimensions in Literature

used for assessment of brand equity, a well-observed challenge is that they are costly to obtain and have a certain lag (i.e., they are not real-time measures). To overcome this, researchers from the Information Systems (IS) discipline have recently alluded to the need to assess and analyze data originating from Social Media Networks (SMN) [16, 17, 18]. Although several frameworks revealing the need to apply SMA techniques have been proposed [19, 20], few manage to provide methodologies for evaluating precise marketing constructs [21, 22, 23, 24] and focus mainly on the financial aspect of brand equity or firm equity.

Marketing researchers [25, 26] have recently stressed the need to include consumer perceptions from SMN data when faced with assessing marketing actions in terms of brand equity. Although similar analyses of SMN data are relatively rare in literature, it is evident that there is a need for an assessment model that shifts from traditional data collection, focuses on the consumer perspective from means such as SMN, and utilizes state of the art computational techniques.

The novelty of this paper lies in the derivation of a computational model that facilitates the assessment of two core dimensions of brand equity; namely brand awareness and brand meaning. Specifically, we assist this process by developing a set of algorithmic tools that are associated with a series of hard computational tasks that have been extensively researched in the past. This paper builds on existing computer science literature that spans a wide range of relevant subfields in order to address the following text and data mining problems: (a) text normalization, (b) corpus vectorization, (c) sentiment analysis, (d) topic modeling and (e) document clustering.

Text normalization [27, 28] refers to the process of removing out of scope information from a given document so that it can be subsequently submitted to the corpus vectorization and topic modeling preprocessing modules. Corpus vectorization, on the other hand [29], relates to the transformation of a large dataset into a set of algorithmically tractable vectors that may, in turn, be utilized as features by the sentiment analysis module. Sentiment analysis [30, 31, 32], in particular, involves training a state of the art machine learning algorithm into classifying a set of pre-labeled positive and negative tweets into the corresponding category by minimizing the associated misclassification cost. Topic modeling [33, 34] focuses on identifying the prevailing topics of discussion in a given corpus. Finally, document clustering [35, 36] addresses the problem of grouping together semantically similar documents which is an imperative prerequisite for estimating the average sentiment value per topic discussed.

Our proposed model draws on Aaker's definition on brand equity [37], Melville's framework [16] for assessing marketing constructs through SNA techniques and utilizes Berry's conception [3] for service brands, using the dimensions of brand awareness and brand meaning. In the next sections we describe the design and evaluation of a computational model that lays out a proposed method on how fundamental dimensions of brand equity can be extracted by mining consumer perceptions from SMN, the computational methods necessary to be followed, and concludes with how these results can be interpreted from a marketing perspective, through the application of the model in a relevant case study.

## 3. Methodology

This study follows a design research approach [38, 39] and positions the work according to Gregor & Hevner's framework [40] following the publication schema as showcased in a paper by McLaren et al. [41]. While brand equity assessment methods are much researched and have been used for many years, it is apparent that there is now a great demand for methods which model the process by including big data techniques in social media data streams. The model falls under the improvement quadrant in the DSR knowledge contribution framework [40], as its main aim is to show how and why this new solution differs from previously presented ones.

Design, implementation and evaluation of the model are justified using prior theory and the case study findings. Our model is grounded in theory for the steps that involve machine learning algorithms, and backed up by empirical evidence for the steps that involve marketing decisions during the assessment life-cycle.

The proposed model is based on two disciplines: marketing and computer science. Each building block includes a number of elements which are illustrated in Figure 1. In this section we introduce the computational model and describe each step in detail.

**Step 1: Prerequisites.** Defining specific keywords that describe the brand and characterize the campaign executed are of utmost importance. The more specific the keywords, the less "noise" the data will generate. Organizations should provide keywords related to the brand (i.e. the product name or company name in the case of services) as well as keywords that de-

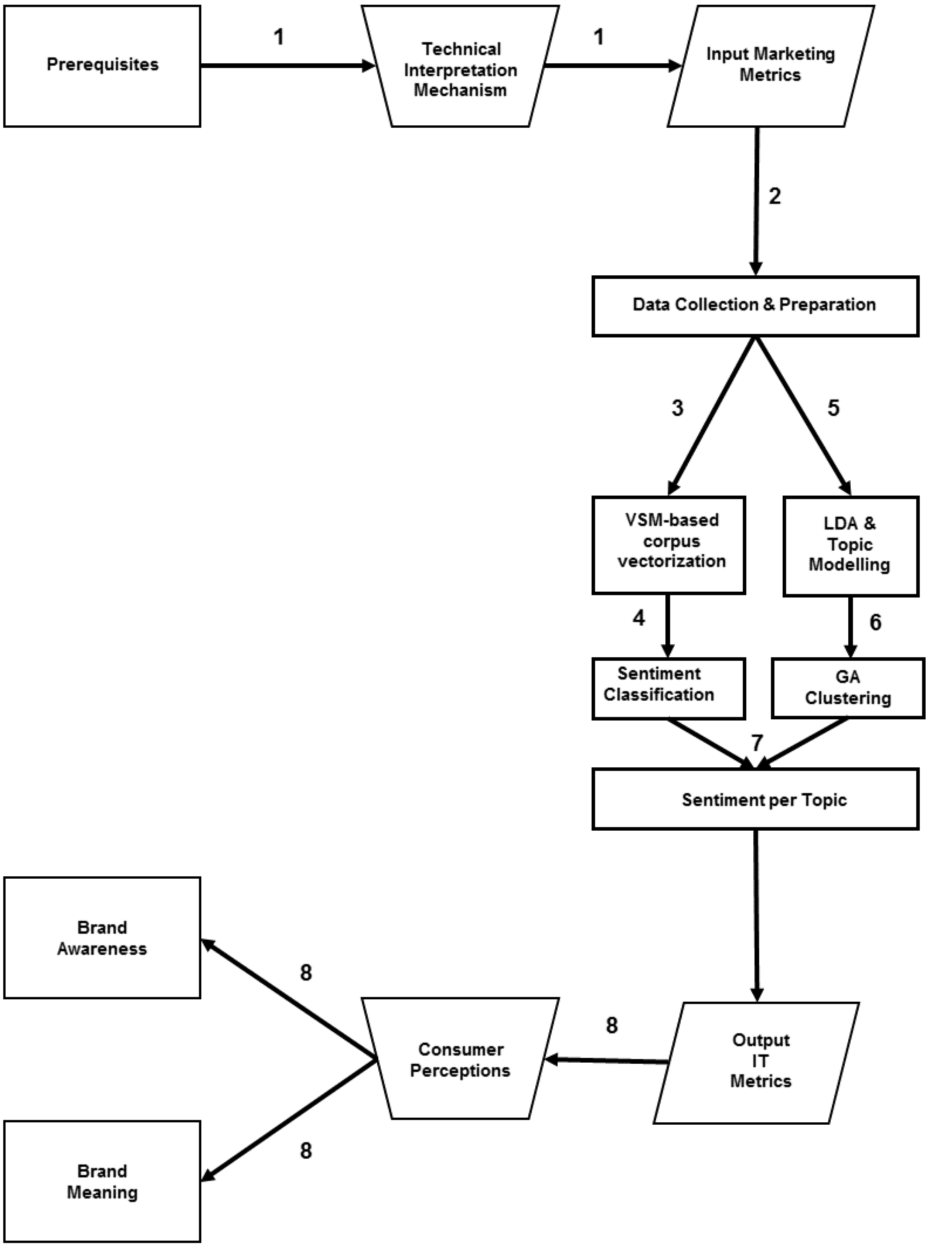

Figure 1: The model

scribe the campaign executed or population targeted (in the case of specific campaign appraisal). These keywords are used during the data collection and preparation steps in order to form a correct corpus of the objective data set.

**Step 2: Data Collection & Preparation.** Defining the social media channel to monitor and extract data is the primary decision that the organization should take. Literature on social media [42, 43] lists pros and cons of each medium and the choice should be based on the desired outcome the organization wishes to obtain. Data are collected through the utilization of the relevant API for the given social media channel and the time period chosen. Finally, once data have been collected, they should subsequently be submitted to a series of data clearing and pre-processing operations. The data preparation process for the subsequent sentiment analysis and topic modeling tasks involves text tokenization into words, elimination of stop words, and stem extraction from each word. Therefore, the final version of the corpus will be formed as a set of purified documents where each document $d \in D$ is a collection of words from a single tweet, given by the following equation:

$$\mathcal{D} = \{d_1, d_2, \ldots, d_n\} \tag{1}$$

where $n$ is the number of available documents.

**Step 3: VSM-based Corpus Vectorization.** The primary objective of this step is to convert each unstructured document into a feature vector that can be subsequently fed into a machine learning algorithm for sentiment classification. Such a transformation aims to obtain a mathematical representation for the corpus so that each document can be treated as a point in a multi-dimensional vector space. A natural approach towards this end is the employment of the standard Vector Space Model (VSM). The main idea behind VSM is to transform each document into a vector, containing only the words that belong to the document and their frequency, by utilizing the so-called “bag of words” representation.

The underlying mathematical abstraction imposed by VSM entails a mapping which transforms the original purified document to its corresponding bag of words representation. That is, each document $d \in \mathcal{D}$ will finally be represented through the utilization of an $M$-terms dictionary

$$T = \{t_1, t_2, \ldots, t_M\} \tag{2}$$

extracted from the purified corpus. This transformation can be formulated by the following equation:

$$\phi : \mathcal{D} \to \mathbb{R}^M \tag{3}$$

such that:

$$\phi(d) = [tf(t_1, d), tf(t_2, d), \ldots, tf(t_M, d)] \tag{4}$$

where $tf(t_i, d)$ is the normalized frequency of term $t_i$ in document $d \in \mathcal{D}$, according to the term frequency-inverse term frequency weighting scheme (TF-IDF). The TF-IDF weighting scheme should be utilized in order to mitigate the effect relating to the complete loss of context information around a term. In this context, each term $t_i$ is assigned a weight $w_i$ of the following form:

$$w_i = idf(t_i, \mathcal{D}) = \log \frac{n}{|\{d \in \mathcal{D} : t_i \in d\}|} \tag{5}$$

In other words, the exploitation of VSM provides a formal representation of the corpus by transforming each set of documents into a corresponding set of feature vectors according to the following equation:

$$\Phi = \phi(\mathcal{D}) = [\phi_1, \phi_2, \ldots, \phi_n] \tag{6}$$

where $\phi_j = \phi(d_j) \in \mathbb{R}^M, \forall j \in [n]$.

**Step 4: Sentiment Classification.** This step encompasses the sentiment classification process which can be further divided into the corresponding training and testing stages. Our approach conducts sentiment analysis through the utilization of a state of the art machine learning algorithm, namely Support Vector Machines (SVMs). SVMs are non-linear classifiers operating in higher-dimensional vector spaces than the original feature space of a given dataset. Their training process involves a quadratic minimization problem, in the context of a binary classification task, which results in a set of optimal parameters, formulated as:

$$\Lambda = \{\lambda_0^*, \lambda_1^*, \ldots, \lambda_m^*\} \tag{7}$$

given that $\lambda_j^* \geq 0$, $\forall j \in [m]$ where $m$ is the number of pre-labeled documents pertaining to the training dataset. The set of optimal parameters $\Lambda$ and the associated training feature vectors define a hyperplane within the implicitly induced higher-dimensional feature space, which serves as the discrimination boundary between the subspaces of positive and negative tweets defined as:

$$g(\phi) = \sum_{j=1}^{m} \lambda_j^* \cdot K(\phi, \phi_j) + \lambda_0^* = 0 \tag{8}$$

such that $-1 \leq g(\phi) \leq +1$ where $K(\cdot, \cdot)$ is the Gaussian kernel function given by the following equation:

$$K(\phi, \phi_j) = \exp(-\gamma \cdot ||\phi - \phi_j||^2) \tag{9}$$

that is employed in order to map the input (TF-IDF)-based feature space into a higher-dimensional vector space. In other words, the discrimination function defined in Equation 8 defines a mapping of the following form:

$$g : \mathbb{R}^M \to [-1, +1] \tag{10}$$

quantifying the distance from the decision boundary, which quantifies the amount of certainty according to which a given tweet is classified as positive or negative. Therefore, decision values close to zero may be indicative of tweets pertaining to the neutral sentiment class. However, such patterns are not explicitly presented to the SVM classifier during training.

The training stage is an essential part of the method, since the application of SVMs on such a large amount of text requires a reasonable amount of labeled data (i.e. texts already classified as positive or negative, based on a marketing perspective classification). This ensures that the SVM algorithm runs with accuracy, providing robust results that limit the amount of fault. These labeled data are in turn used by the SVM algorithm as a benchmark, in order to score the number of texts that are in scope of the sentiment exercise. The testing stage, on the contrary, aims at testing the accuracy and validity of the SVM algorithm on the largest subset of the dataset that was not previously classified. The sentiment classification module assigns each document's $d_j$ feature vector $\phi_j$ with a soft decision value $s_j = g(\phi_j)$ which is indicative of the positive or negative sentiment strength. Finally, the overall output of this step is a set $S$ of sentiment values given by:

$$S = \{s_1, s_2, \ldots, s_n\} \tag{11}$$

such that $-1 \leq s_j \leq +1,\ \forall j \in [n]$. The soft decision values appearing in Equation 11 are in fact the output of the trained SVM classifier during the testing stage. This is subsequently utilized in order to estimate the average sentiment value for a given subset of tweets.

**Step 5: LDA - Topic Modeling.** This step commits to the LDA probabilistic topic modeling algorithm, which besides unraveling the latent topic structure of the corpus, lays the foundations for an alternative vectorized corpus representation. Probabilistic topic modeling approaches share the fundamental assumption that documents within a corpus can be formulated as mixtures of topics, where each topic is modeled as a probability distribution over words. Therefore, a topic model may be interpreted as a generative model for documents, since it specifies a simple probabilistic procedure according to which new documents emerge.

In this context, each document $d_j \in \mathcal{D}$ may be treated as a point into a $T$-dimensional probability vector space $\mathcal{P} = [0,1]^T$. Formally, the application of LDA defines a mapping of the following form:

$$\psi : \mathcal{D} \rightarrow \mathcal{P} \tag{12}$$

where each document $d_j \in D$ is mapped to a point $\psi_j = \psi(d_j) \in [0,1]^T$, such that:

$$\sum_{t=1}^{T} \psi_j(t) = 1, \ \forall j \in [n] \tag{13}$$

The previous definitions imply that the set of documents $\mathcal{D}$ acquires an alternative vector representation according to the following equation:

$$\Psi = \psi(\mathcal{D}) = \{\psi_1, \psi_2, \ldots, \psi_n\} \tag{14}$$

**Step 6: GA Clustering.** The fundamental challenge addressed by this step is to organize the given corpus into a predefined number of $K$ semantically coherent and highly interpretable groups of documents. The semantic coherence directive may be achieved by grouping together documents whose LDA-based vector representations exhibit minimum topic deviation with respect to the corresponding cluster centroids. The existence of highly interpretable groups of documents is, in turn, associated with cluster centroids that tend to accumulate the majority of their probability mass on a single topic.

In this paper, we devise a novel evolutionary clustering mechanism which relies on a centroid-based encoding scheme of the possible clustering solutions. This encoding scheme provides a significant improvement in handling the inherent NP-completeness of the underlying clustering problem, especially for instances of vast data volumes, since the proposed genetic encoding does not depend on the size of the data set. The novelty of our genetic clustering method relates to the fact that the constituent initialization, mutation and crossover operators take into significant consideration the particularity of the underlying search space. Specifically, our genetic clustering algorithm is mediated by a set of genetic operators that function exclusively within the $T$-dimensional standard simplex. Moreover, the proposed initialization operator is particularly designed so that the underlying evolutionary search procedure commences within the close neighborhood of semantically focused cluster centroids.

Formally, given the LDA-based representation of our corpus $\Psi$, our evolutionary clustering method consists in determining a set $\Psi^*$ of $K$ cluster centroids, given as:

$$\Psi^* = \{\Psi_1^*, \ldots, \Psi_K^*\} \tag{15}$$

that implicitly define an optimal $K$-partitioning of $\Psi$ such that:

$$\Psi = \bigcup_{i=1}^{K} \Psi_i \tag{16}$$

and

$$\Psi_r \cap \Psi_l = \emptyset,\ \forall r, l \in [K] : r \neq l \tag{17}$$

where

$$\Psi_i = \{\psi \in \Psi : ||\psi - \Psi_i^*|| < ||\psi - \Psi_r^*||,\ \forall r \in [K] : r \neq i\} \tag{18}$$

Optimality of the $K$-partitioning is measured in terms of the aggregated topic deviation around the corresponding cluster centroids and is enforced by addressing the following optimization problem:

$$\min_{\{\Psi_1^*, \ldots, \Psi_K^*\} \in \mathcal{P}^K} F_{topic\_deviation}(\Psi^*, \Psi) \tag{19}$$

where the objective function to be minimized by the proposed centroid-based genetic algorithm is given by the following equation:

$$F_{topic\_deviation}(\Psi^*, \Psi) = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{|\Psi_i|}\sum_{r=1}^{|\Psi_i|}||\Psi_i^r - \Psi_i^*||^2 \tag{20}$$

where $\Psi_i^r$ is the $r$-th LDA-based feature vector pertaining to the $i$-th cluster. Minimization of the objective function defined in Equation 20 within the context of genetic algorithms can be achieved through the definition of appropriate initialization, crossover and mutation operators that guarantee the efficient exploration of the underlying problem space.

The most important feature of the proposed population initialization routine is that all initial solutions comply with the underlying constraint that requires all cluster centers to lie within the $T$-dimensional simplex. In this context, initial clustering solutions are generated within two groups. The first group contains solutions that are located on the corners of the $T$-dimensional simplex that is the unit basis vectors of $\mathcal{P}$. The second group of solutions is generated by combining pairs of corner-located solutions. Letting $C_a$ and $C_b$ be two corner located solutions, a clustering solution $C_r$ pertaining to the second group will be formed as a point within the line segment defined by the points $C_a$ and $C_b$ as:

$$C_r = C_a \cdot R + C_b \cdot (1 - R) \tag{21}$$

where R is uniformly sampled within the $(0, 1)$ interval, keeping in mind that the dimensionality of the standard simplex is defined by the number of topics of the LDA model.

The proposed crossover operator takes into account both the centroid-based representation of solutions and the underlying linear constraints that must be satisfied. These linear constraints take into consideration the fact that all cluster centers must lie within the $T$-dimensional standard simplex. The crossover operations involve randomly selecting a number of crossover pairs corresponding to the parents that will contribute to the generation of a single crossover child. The rationale behind the adopted crossover operation builds upon the generation of random points within the line segment, defined by the pair of the selected crossover parents according to Equation 21.

The mutation operations performed by the centroid-based genetic algorithm take into account both the centroid-based representation of solutions and the underlying linear constraints that must be satisfied. These linear constraints once again take into consideration the fact that all cluster centers must lie within the $T$-dimensional standard simplex. The mutated offspring are generated by utilizing two major operations. The first operation involves determining the best solution within the parent population. Subsequently, by retrieving the corresponding cluster centers, each cluster center is moved towards the center of the line segment connecting the nearest and the furthest point within the dataset. For each cluster center a number of additional cluster centers will be incorporated within the mutated population. The second operation involves randomly selecting a number of coordinate-indices for each cluster center and subsequently performing a random permutation of the selected cluster-center coordinates. This procedure guarantees that the mutant offspring will also lie within the $T$-dimensional standard simplex.

**Step 7: Sentiment per Topic.** The ultimate purpose of this step is to assign each cluster and associated prevailing topic with an average sentiment value. This task may be achieved by firstly considering the set of documents that pertain to a given cluster designated as:

$$\Psi_i = \{\Psi_i^1, \cdots, \Psi_i^{n_i}\} \tag{22}$$

where $n_i$ is the number of documents forming the $i$-th cluster. The same grouping principle may be applied to the corresponding set $S$ of sentiment values such that

$$S_i = \{S_i^1, \cdots, S_i^{n_i}\} \tag{23}$$

provides the sentiment value assigned to each document of the $i$-th cluster. Therefore, the average sentiment value may be easily computed as:

$$S_i^* = \frac{1}{n_i} \sum_{r=1}^{n_i} S_i^r \tag{24}$$

This average sentiment value, in particular, may be assigned to the corresponding prevailing topic $T_i^*$ which is the one accumulating the majority of the probability mass according to the following equation:

$$T_i^* = \arg\max_{t \in T} \Psi_i^*(t) \tag{25}$$

**Step 8: Output IT Metrics & Consumer Perceptions.** The model generates prevailing topics and clusters in the given corpus, while also producing four key output metrics:

- **Metric no1:** Volume of Tweets / per time period
- **Metric no2:** Sentiment Classification / per time period
- **Metric no3:** Volume per topic / per time period
- **Metric no4:** Sentiment per topic / per time period

Metric pairs 1 & 3 and 2 & 4 provide insights on brand awareness and brand meaning respectively.

## 4. Case Study & Results

Uber Technologies Inc. is an American international transportation network company headquartered in San Francisco, California, founded by Travis Kalanick and Garrett Camp in 2009. The company develops, markets and operates a mobile app which gives the ability to smartphone users to search for and request a trip pickup from an exact geographical location, from a designated Uber driver. As of 28 May, 2015, the service was available in 58 countries and 300 cities worldwide[1].

One of the official channels for communication with Uber, as listed in their website, is Uber's Twitter account (@UBER). On a 24/7 basis, Uber is the subject of on-going requests through their Twitter account, that usually relate to, but are certainly not limited to, pricing issues, service complaints, reporting of dangerous driving, user privacy and safety issues.

**Step 1: Prerequisites.** To illustrate the validity of our model, we unveil consumer perceptions relating to the Uber brand. The data collection process was focused on gathering tweets that were explicitly referring to the Uber transportation network by performing hashtag and mention filtering on the terms "#uber" and "@uber".

**Step 2: Data Collection & Preparation.** We collected and analyzed a set of over 280.000 tweets during a three-month period, between January 2015 and April 2015, by utilizing the Streaming API of Twitter. Data were collected on a 24/7 basis by parsing Twitter's Streaming API and stored in a dedicated MySQL database server. Appropriate Python code and Linux wrappers ensured stability and recovery in case of network downtime. Data preparation involved the elimination of all non-English tweets, non-letter characters, URLs, mentions and re-tweet identifiers. Therefore, our corpus was constructed as a collection of distinct author documents where each document contained the purified text from a single tweet. The final version of the corpus was formed after applying a series of tokenization and stop word removal on the original tweet text. Moreover, we deleted all words whose length was less than two characters. The final version of the corpus

[1] http://www.uber.com. Last Accessed 22/08/2016

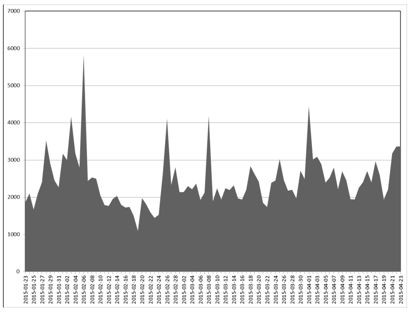

Figure 2: Daily Volume

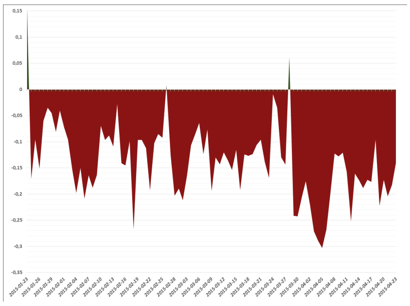

Figure 3: Daily Sentiment

involved 221.958 tweets. Figure 2 depicts the evolution of the daily volume of tweets gathered.

**Step 3: VSM-based Corpus Vectorization.** Each document of the corpus was transformed into a vector, containing only the words that belong to the document and their frequency, by utilizing the so-called “bag of words” representation. The main idea behind this exercise was to represent each document exclusively by the words it contains by tokenizing sentences into elementary term (word) elements, and losing the associated punctuation, order and grammar information. The size of the underlying dictionary of terms defined in Equation 2 was experimentally set to $M = 400$ in order to avoid sparse VSM representations where only a small fraction of the resulting feature vectors have non-zero elements.

**Step 4: Sentiment Classification.** Sentiment classification was performed in a binary classification setting by utilizing the Gaussian kernel function defined in Equation 9 where the parameter $\gamma$ was experimentally set to 1. Performance evaluation of the SVM classifier was measured by adopting the standard 10-fold cross-validation process on an equally balanced set of 2.164 previously labeled Tweets. Each fold involved splitting the complete set of pre-labeled samples into a 95% training data / 5% testing data ratio, where the first subset of data instances was utilized to build the classifier and the latter for assessing its ability to infer the sentiment polarity of unseen data patterns. The training classification efficiency-related measurements are summarized in Table A.5. The sentiment categorization for the rest of the un-labeled data patterns was conducted by exploiting the complete set of pre-labeled data instances so that the trained classifier accumulated the maximum amount of available knowledge for the problem of sentiment classification. At this stage each tweet in our corpus was assigned a unique soft decision value within the $[-1, +1]$ interval indicating the amplitude of negative or positive sentiment. Figure 3 indicatively depicts the average sentiment of tweets per day for the given data range period (green positive, red  negative).

**Step 5: LDA - Topic Modeling.** LDA Topic Modeling was conducted by setting $T = 10$, aiming at retrieving the ten most discussed topics in our corpus of 221.958 tweets. This particular decision lies upon our intention to generate a more abstract overview of the prevailing topics that users chose to discuss. We have also experimented by varying $T$ within the discrete $\{5, 10, 20, 30, 40, 50, 100\}$ interval. Our experiments verified that for values of $T$ that are greater than 50 the majority of the documents acquire a zero-valued or totally uniform vector representation of the associated probability distribution. That is, for significantly large values of $T$ most of the documents in our corpus either fail to be represented or they are evenly distributed within the underlying semantic space. Both cases render the subsequent topic clustering step infeasible. In the context of our work, however, the value of $T$ depends heavily on the amount of semantic granularity that is sought to be achieved. It is extremely important to mention that this particular pre-processing step is, in fact, coupled with the subsequent semantically-coherent organization task of our corpus. Therefore, the most important factor relates to choosing the number of clusters as equal to the number of topics. Detailed results on the ten most discussed topics are depicted in Table 2.

| Topic Number | Topic Theme |
|---|---|
| Topic 1 | Uber service |
| Topic 2 | Uber as a start-up |
| Topic 3 | Coupons |
| Topic 4 | Innovation |
| Topic 5 | Free codes |
| Topic 6 | Support - Help |
| Topic 7 | Selfies in back of Uber cars |
| Topic 8 | Surge Pricing |
| Topic 9 | Women & Uber |
| Topic 10 | #Ubered |

Table 2: Topics

| No of Clusters | Genetic Algorithm | K-Means Algorithm |
|---|---|---|
| **2** | **0,332784** | 0,341310 |
| **3** | **0,242544** | 0,275912 |
| **4** | **0,209119** | 0,246677 |
| **5** | **0,188783** | 0,235065 |
| **6** | **0,178833** | 0,206715 |
| **7** | **0,158493** | 0,198441 |
| **8** | **0,162271** | 0,166213 |
| **9** | **0,150122** | 0,193225 |
| **10** | **0,118459** | 0,166359 |

Table 3: GA vs. K-Means

| Cluster | Topic | Probability Mass | No. Tweets |
|---|---|---|---|
| 1 | Surge Pricing | 0,6620 | 22.346 |
| 2 | Innovation | 0,6814 | 21.743 |
| 3 | Women & Uber | 0,7137 | 28.403 |
| 4 | Selfies in back of Uber cars | 0,6829 | 16.488 |
| 5 | Uber service | 0,7201 | 25.734 |
| 6 | # Ubered | 0,6592 | 28.718 |
| 7 | Coupons | 0,6696 | 17.434 |
| 8 | Free codes | 0,6682 | 22.964 |
| 9 | Uber as a start-up | 0,6706 | 17.463 |
| 10 | Support - Help | 0,6502 | 20.665 |

Table 4: Prevailing topic of each cluster — No. of Tweets

**Step 6: GA Clustering.** At this point all tweets are represented as probabilistic mixtures of the ten prevailing topics in the given corpus and are assigned a particular sentiment value. The challenge addressed here is to cluster the complete set of tweets in a predefined number $K$ of semantically-focused and minimally topic-deviated clusters, enabling us to reveal the average sentiment value associated with a certain topic. To this end we applied our full corpus of tweets through the application of the previously introduced GA, starting iteratively from $K = 2$ to $K = 10$, in order to determine the number of clusters and corresponding cluster centroids that induce the overall minimum topic deviation. Results of the GA clustering topic deviation in comparison to $K$-means are depicted in Table 3, indicating that the best clustering configuration is the one obtained for the number of $K = 10$ clusters for the centroid-based GA. The fact that the overall minimum of the topic deviation measure is achieved for the number of ten clusters provides significant justification towards inferring that the true number of clusters for this particular dataset is also ten. This, in turn, is no coincidence since $T = 10$ is the dimensionality of the LDA-induced semantic space that underlies the given corpus. The GAbased clustering results indicate that each group of tweets is minimally distributed around the corresponding cluster centroids, where each cluster centroid accumulates the vast majority of the probability mass on a single topic as presented in Table 4.

**Step 7: Sentiment per Topic.** The final step involves sentiment classification of tweets participating in each cluster, depicted at a daily average as per Figure 5.

**Step 8: Output IT Metrics & Consumer Perceptions.**

Interpretation of the output metrics is discussed in the next section.

## 5. Discussion

Application of our computational model in this vast pool of “Uber”-related data has generated significant insights, which are of particular interest when assessing the given brand. Starting from a collection of more than

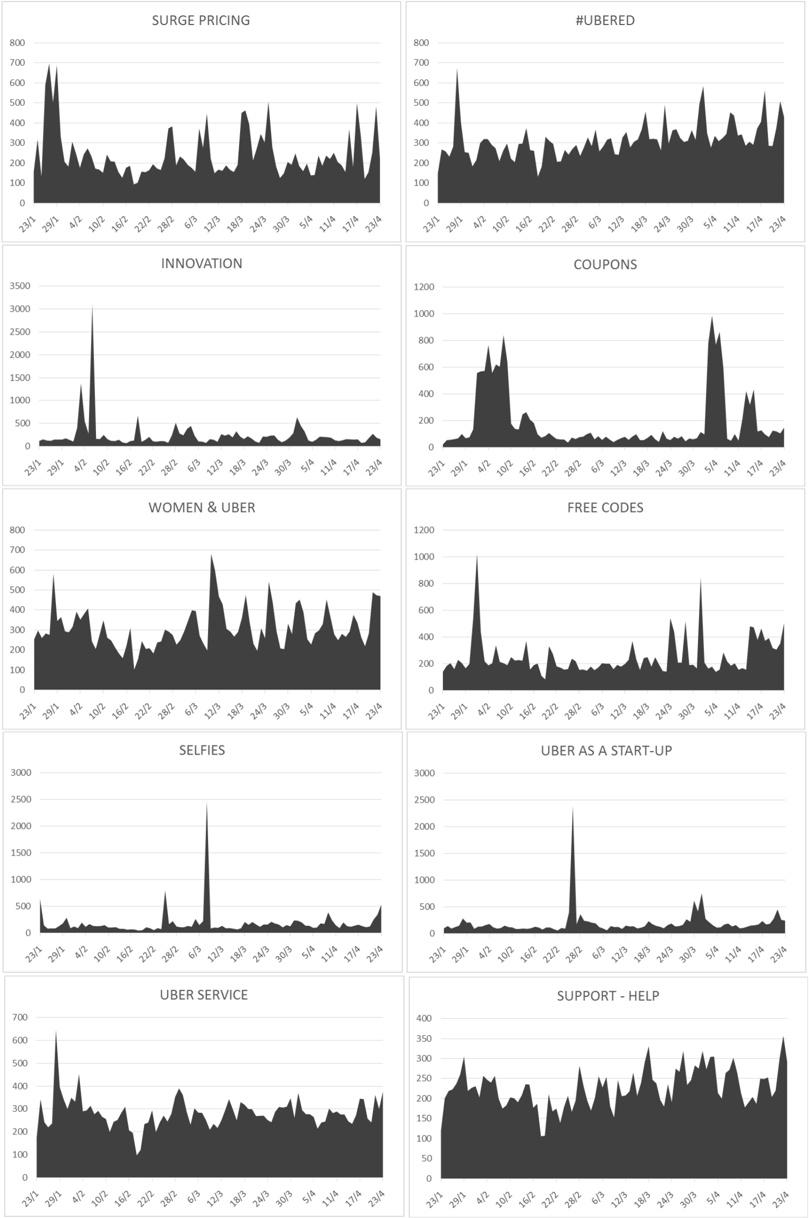

Figure 4: Volume per Cluster

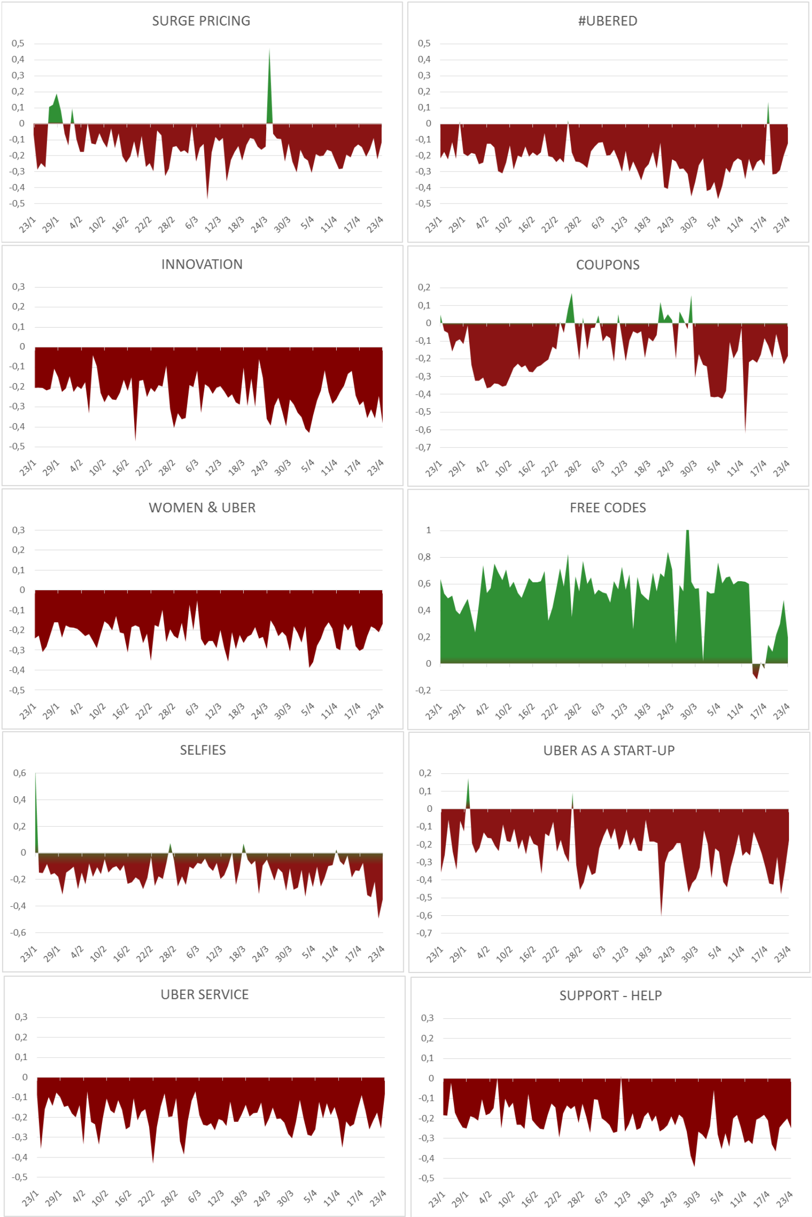

Figure 5: Sentiment per Cluster

280.000 tweets, the model manages to generate four key insights, with regard to (a) overall volume of tweets and prevailing topics discussed, (b) daily sentiment towards the brand, (c) optimal clustering of tweets in semantically-coherent clusters under a distinct topic, and (d) an overall average sentiment assessment of each topic per day.

Figures 2 and 3 aptly depict the average volume of tweets and daily sentiment of users towards Uber. The overall volume of tweets is evenly distributed at an average of 2.500 tweets per day. The overall sentiment for the given time period indicates moderately negative sentiment towards the brand, with an exception on the date range between 27-28/03. This metric provides an overall view of customer perception towards the brand but does not explain the key reasons behind this polarity. The next step is to discover prevailing topics that were discussed by users in the data corpus. Table 4 and Figure 4 summarize this information for the given time period. A random selection of tweets that indicatively showcase users' posts per topic can also be found in the Appendix (Table A.6). Our data analysis reveals that users pay particular interest, when choosing to express their opinions via Twitter, to the following subjects:

- **Uber service:** A vast majority of users chose to directly contact Uber by preceding their tweets with the "@" symbol and comment on their experience with Uber. Tweets indicate complaints in relation to drivers, dangerous driving, dirty fleet, and incidents of overcharging.
- **Uber & start-ups:** This particular topic revealed coverage by Twitter media accounts and tech aficionados, commenting on how Uber is a start-up that has disrupted the tech and transport scene and how start-ups wishing to disrupt the market refer to their service as "Uber for the specific market"

- **Coupons:** High number of tweets promoting coupons for reduced fares. Usually re-tweeted by an increased number of accounts.
- **Innovation:** A particular discussion about Uber launching a research lab in the US, for self-driving cars, which gained high interest from Twitter news media accounts.
- **Free codes:** Users expressing their gratitude for free codes that resulted in free rides.
- **Support  Help:** Users using Twitter as the primary means of contact for support and help issues related to theft, sexual abuse, general complaints and cancellation of service.
- **Selfies in back of Uber cars:** Users posting selfies with specific hashtags in the back of Uber cars and celebrities which were re-tweeted posting their selfies.
- **Surge Pricing:** Huge interest and discussion about Uber's surge pricing algorithm which in many cases frustrates users due to overcharging.
- **Women & Uber:** Specific discussion sparked by Uber's initiative to create jobs for women.
- **#Ubered:** Trending hashtag applied when having a bad experience or worse on Uber. This type of expression has become a meme between Uber users especially when resulting from paying fares way above the normal taxi rate.

At this point all tweets hold a sentiment value, are part of a unique cluster and fall under a single topic. The final step in order to unveil the key factors that drove customer sentiment towards the brand is the extraction of sentiment per cluster. The metric depicted in Figure 5 indicates users' perceptions towards the brand for a specific topic, weighted against a sentiment scale, indicating brand meaning and value towards the respective features of the brand. Results reveal that Twitter users are particularly negative towards Uber's service, support, and strategy of expansion in the market.

On the contrary, users fairly seem to enjoy the promotion in terms of free riding codes that Uber offers.

## 6. Conclusion & Future Work

We presented a computational framework that combines topic modeling, data clustering and sentiment analysis, which is part of an overall brand equity assessment model based on big data and machine learning. We empirically applied the computational framework to guide the analysis of over 280.000 tweets in relation to the Uber transportation network and users who included the specific brand as part of their tweeting activity. Our results indicate that the dataset is inherently organized in ten semantically-focused clusters, each one minimally distributed around a unique topic. A direct consequence of this fact is that we can associate each cluster and corresponding topic with an average sentiment value, thus unraveling the public attitude against particular aspects of the brand under investigation.

Our research contributes to the literature in an interdisciplinary scope, drawing upon big data and machine learning techniques for improved data analysis, as well as marketing, by presenting a novel method for assessing specific brand equity dimensions through mining consumer perceptions in SMN. From a machine learning perspective, we contend that the obtained topic clustering results indicate significant improvement in extracting semantically-focused groups of documents, when compared against traditional clustering algorithms such as the k-means. The clustering superiority of our proposed genetic algorithm is also justified by measuring the intra- and inter-cluster semantic distances of the obtained cluster formations. From a marketing perspective, we stress the need to complement traditional methods of brand equity assessment with a combined marketing-driven approach paired with big data and machine learning techniques.

This study also presents some insightful managerial implications. Marketing analysts may use the results generated from the proposed framework to uncover sentiment tendencies as well as prevailing topics and meaning that drove discussion towards their brand. For example, our analysis revealed that Twitter users are particularly negative towards Uber's service, support, and strategy of expansion in the market. On the contrary, users fairly seem to enjoy the promotion in terms of free riding codes that Uber offers. Such knowledge may be used to drive future corporate actions and decisions in order to further strengthen the positive aspects of the corporate image as this is formulated through discussions in online social networks.

Our study has several limitations that can act as incitement for future research. First, the case study is limited to a single social media channel (i.e. Twitter). The proposed framework is designed in such a way that can be applicable to any medium that provides the relevant API for data extraction. Behavioral aspects of tweeting activity such as hearts, re-tweets without comment and meta attributes such as number of followers, geo-location, etc., are an additional pool of information that could further enhance similar models. Our model explicitly focuses on the semantic attributes of the tweets (text analysis) which capture comments and re-tweets that include a comment, as the core of our computational model leverages machine learning techniques that handle data in text form. This information could and should be considered by future researchers that wish to expand on our model.

A second limitation is that results derived from mining any online medium for sentiment don't necessarily align with real world viewpoints, mainly due to the characteristics of the population choosing to express views from the specific mean. Although positive attributes of customer perceptions may be extracted by various methods, negative attributes will primarily be reflected from consumer perceptions generated from social media, due to the nature of the medium that embraces direct communication. Another limitation is the number of topics and clusters the researcher will choose to set when applying the model. Slicing and dicing of data should be at the discretion of the researcher and multiple iterations of the approach might be needed to reach optimal level of detail results. Lastly, in order to assure generalizability of the results, the model should be applied in organizations of different verticals and market sizes. We urge researchers wishing to apply our proposed model, to take these parameters into account in the experimental applications they choose to proceed with.

## Acknowledgments

This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program “Education and Lifelong Learning” of the National Strategic Reference Framework (NSRF) - Research Funding Program: Aristeia, SocioMine.

## References

[1] C. Duan, C. E. Hill, The current state of empathy research, Journal of Counseling Psychology 43 (3) (1996) 261.

[2] B. Yoo, N. Donthu, Developing and validating a multidimensional consumer-based brand equity scale, Journal of Business Research 52 (1) (2001) 1–14.

[3] L. L. Berry, Cultivating service brand equity, Journal of the Academy of Marketing Science 28 (1) (2000) 128–137.

[4] D. A. Aaker, The value of brand equity, Journal of Business Strategy 13 (4) (1992) 27–32.

[5] K. L. Keller, Conceptualizing, measuring, and managing customer-based brand equity, Journal of Marketing 57 (1) (1993) 1–22.

[6] G. Christodoulides, L. de Chernatony, Consumer-based brand equity conceptualization and measurement: A literature review, International Journal of Market Research 52 (1) (2010) 43–66.

[7] C. J. Cobb-Walgren, C. A. Ruble, N. Donthu, Brand equity, brand preference, and purchase intent, Journal of Advertising 24 (3) (1995) 25–40.

[8] R. Pappu, P. G. Quester, R. W. Cooksey, Consumer-based brand equity: improving the measurement-empirical evidence, Journal of Product & Brand Management 14 (3) (2005) 143–154.

[9] J. H. Washburn, R. E. Plank, Measuring brand equity: An evaluation of a consumer-based brand equity scale, Journal of Marketing Theory and Practice 10 (1) (2002) 46–62.

[10] G. Christodoulides, L. de Chernatony, O. Furrer, E. Shiu, T. Abimbola, Conceptualising and measuring the equity of online brands, Journal of Marketing Management 22 (7-8) (2006) 799–825.

[11] D. M. Blei, Probabilistic topic models, Communications of the ACM 55 (4) (2012) 77–84.

[12] A. French, G. Smith, Measuring brand association strength: a consumer based brand equity approach, European Journal of Marketing 47 (8) (2013) 1356–1367.

[13] X. Tong, J. M. Hawley, Measuring customer-based brand equity: empirical evidence from the sportswear market in China, Journal of Product & Brand Management 18 (4) (2009) 262–271.

[14] W. Lassar, B. Mittal, A. Sharma, Measuring customer-based brand

equity, Journal of Consumer Marketing 12 (4) (1995) 11–19.

[15] C. Veloutsou, G. Christodoulides, L. de Chernatony, A taxonomy of measures for consumer-based brand equity: drawing on the views of managers in Europe, Journal of Product & Brand Management 22 (3) (2013) 238–248.

[16] P. Melville, V. Sindhwani, R. Lawrence, Social media analytics: Channeling the power of the blogosphere for marketing insight, Proc. of the WIN 1 (1) (2009) 1–5.

[17] G. C. Kane, M. Alavi, G. Labianca, S. P. Borgatti, What's different about social media networks? a framework and research agenda, MIS Quarterly 38 (1) (2014) 275–304.

[18] Z. Dongsong, Y. Wei Thoo, Social media use in decision making: Special issue of decision support systems for the 10th workshop on e-business, Decision Support Systems 63 (2014) 65–66.

[19] T.-P. Liang, E. Turban, Introduction to the special issue social commerce: a research framework for social commerce, International Journal of Electronic Commerce 16 (2) (2011) 5–14.

[20] W. Fan, M. D. Gordon, The power of social media analytics, Communications of the ACM 57 (6) (2014) 74–81.

[21] A. H. Zadeh, R. Sharda, Modeling brand post popularity dynamics in online social networks, Decision Support Systems 65 (2014) 59–68.

[22] L. Callarisa, J. S. García, J. Cardiff, A. Roshchina, Harnessing social media platforms to measure customer-based hotel brand equity, Tourism Management Perspectives 4 (2012) 73–79.

[23] X. Luo, J. Zhang, W. Duan, Social media and firm equity value, Information Systems Research 24 (1) (2013) 146–163.

[24] Y. Yu, W. Duan, Q. Cao, The impact of social and conventional media on firm equity value: A sentiment analysis approach, Decision Support

Systems 55 (4) (2013) 919–926.

[25] K. S. Coulter, M. Bruhn, V. Schoenmueller, D. B. Schäfer, Are social media replacing traditional media in terms of brand equity creation?, Management Research Review 35 (9) (2012) 770–790.

[26] A. Culotta, J. Cutler, Mining brand perceptions from Twitter social networks, Marketing Science 35 (3) (2016) 343–362.

[27] E. Clark, K. Araki, Text normalization in social media: progress, problems and applications for a pre-processing system of casual English, Procedia-Social and Behavioral Sciences 27 (2011) 2–11.

[28] A. Mikheev, Document centered approach to text normalization, in: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2000, pp. 136–143.

[29] G. Salton, A. Wong, C.-S. Yang, A vector space model for automatic indexing, Communications of the ACM 18 (11) (1975) 613–620.

[30] N. F. Da Silva, E. R. Hruschka, E. R. Hruschka, Tweet sentiment analysis with classifier ensembles, Decision Support Systems 66 (2014) 170–179.

[31] M. Salehan, D. J. Kim, Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics, Decision Support Systems 81 (2016) 30–40.

[32] E. Fersini, E. Messina, F. A. Pozzi, Sentiment analysis: Bayesian ensemble learning, Decision Support Systems 68 (2014) 26–38.

[33] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 1999, pp. 50–57.

[34] D. Cai, Q. Mei, J. Han, C. Zhai, Modeling hidden topics on document manifold, in: Proceedings of the 17th ACM Conference on Information

and Knowledge Management, ACM, 2008, pp. 911–920.

[35] W. Song, S. C. Park, Genetic algorithm-based text clustering technique, in: International Conference on Natural Computation, Springer, 2006, pp. 779–782.

[36] W. Xu, X. Liu, Y. Gong, Document clustering based on non-negative matrix factorization, in: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2003, pp. 267–273.

[37] D. A. Aaker, Measuring brand equity across products and markets, California Management Review 38 (3) (1996) 102–120.

[38] S. Gregor, D. Jones, The anatomy of a design theory, Journal of the Association for Information Systems 8 (5) (2007) 312.

[39] R. H. Von Alan, S. T. March, J. Park, S. Ram, Design science in information systems research, MIS Quarterly 28 (1) (2004) 75–105.

[40] S. Gregor, A. R. Hevner, Positioning and presenting design science research for maximum impact, MIS Quarterly 37 (2) (2013) 337–355.

[41] T. S. McLaren, M. M. Head, Y. Yuan, Y. E. Chan, A multilevel model for measuring fit between a firm's competitive strategies and information systems capabilities, MIS Quarterly 35 (4) (2011) 909–929.

[42] L. Safko, The social media bible: tactics, tools, and strategies for business success, John Wiley & Sons, 2010.

[43] A. M. Kaplan, M. Haenlein, Users of the world, unite! The challenges and opportunities of social media, Business Horizons 53 (1) (2010) 59–68.

# AppendixA.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Accuracy per fold:** | | | | | | | | | |
| 0,83240223 | 0,83798883 | 0,93854749 | 0,87640449 | 0,96629213 | 0,97752809 | 0,79775281 | 0,86440678 | 0,84745763 | 0,92090395 |
| **Mean Accuracy:** | 0,886 (+/- 0,116) | | | | | | | | |
| **Precision per fold:** | | | | | | | | | |
| 0,85882353 | 0,90789474 | 0,94444444 | 0,93670886 | 1,00 | 0,97802198 | 0,92307692 | 0,875 | 0,83870968 | 0,87254902 |
| **Mean Precision:** | 0,914 (+/- 0,100) | | | | | | | | |
| **Recall per fold:** | | | | | | | | | |
| 0,8021978 | 0,75824176 | 0,93406593 | 0,81318681 | 0,93406593 | 0,97802198 | 0,65934066 | 0,85555556 | 0,86666667 | 0,98888889 |
| **Mean Recall:** | 0,859 (+/- 0,198) | | | | | | | | |
| **F-Score per fold:** | | | | | | | | | |
| 0,82954545 | 0,82634731 | 0,93922652 | 0,87058824 | 0,96590909 | 0,97802198 | 0,76923077 | 0,86516854 | 0,85245902 | 0,92708333 |
| **Mean F-score:** | 0,882 (+/- 0,129) | | | | | | | | |

Table A.5: Sentiment Classification Results

| |
|---|
| **Cluster 1: Surge Pricing** |
| yourfriendandy. (February 2, 2015). @Uber surge rates are the worst. |
| Siddharth_T. (February 23, 2015).,@Uber I hate you guys for exploiting via surge. I take uber all over the world, today Is going to be last day I use Uber. #UberPhoenix |
| ashleysueanna. (February 2, 2015).,I love how during surge pricing my driver takes me the most unnecessary route to get to Philz. @Uber you are the worst #SanFrancisco |
| **Cluster 2: Innovation** |
| U_DRIVERS.,(February 2, 2015). #Uber Opening #Robotics,#Research Facility In Pittsburgh To Build #selfdrivingcar,http://t.co/sayteEUtyo via @techcrunch,@U_DRIVERS |
| DustinStiver. (February 3, 2015). RT @SCSatCMU: CMU and @Uber announce research on mapping, vehicle safety, autonomy, New tech centre in Pittsburgh http://t.co/y5CsXFCxA7 |
| **Cluster 3: Women & UBER** |
| _rfenton.,(February 12, 2015).,#uber,#is #terrible #for #women http://t.co/CscQ3vfDk1 |
| raymondchung. (March 11, 2015).,“#Uber seeks to fix its gender problem with UN partnership and promise to create 1M jobs for women” #GenderEquality http://t.co/jiKC5FbpSG |
| juhasaarinen. (March 11, 2015).,“Uber commits to creating 1,000,000 jobs for women globally on the @Uber platform by 2020.” Pretty exact number. |
| **Cluster 4: Selfies in UBER cars** |
| Popwrecked. (March 7, 2015).,#SexySaturday #SuperModel @AlexandriaMorgz is #PopwreckedApproved (even in the back of an @Uber car)! |
| GinnyMcQueen. (February 2, 2015).,#Uber #selfie #sunglasses #weirdo http://t.co/H3NuSTWsNu |
| **Cluster 5: UBER Service** |
| RANOPLAN. (March 26, 2015). @Uber just had the worst service ever! I paid and the driver drop me off is this a new service ????? |
| lizBpimpin. (March 28, 2015). @Uber might be THE worst ripoff and piece of trash service I have EVER had the misfortune to take. 40 minute wait for a 6 minute ride at 8?! |
| TheGangGreen34. (March 8, 2015) @Uber Worst service EVER!!!,Take a regular cab.,Was charged over $100 for a 15 mile ride that was quoted at $30 and driver was #clueless. |
| **Cluster 6: #UBERed** |
| iLostMyDolphin. (April 4, 2015).,#ubered RT @LaraRanallo: @Uber are you serious?,Your columbus drivers are the worst!,Is there any care about customer service? #uber |
| qmleasing. (March 12, 2015).,When Your 20 Minute @Uber Ride From The Airport Costs More Than Your Airfare, YOU GOT #UBERED http://t.co/HvfYoI2Ydk |
| t4xynatty.,(March 21, 2015).,Was asked directions yesterday by a #Uber,driver as he was lost. Customer was so pissed off got out of his car and into,my TAXI... #ubered |
| **Cluster 7: Coupons** |
| UberRiders. (January 31, 2015). Get $30 Off Your #UBER #RIDE w Uber #Promo #Code “UberComeGetMe” http://t.co/cXTlGloXx5 #GOPATS 17 |
| mcaldwellauthor. (April 16, 2015) RT @HybridVigorFilm: Anyone want a free #UBER ride? Enjoy! Coupon code: x6zlv |
| **Cluster 8: FREE Codes** |
| Gracefulofshit. (February 22, 2015). @Uber thanks for ignoring the free ride promo code I entered and charging me $30 |
| PMPEire. (April 9, 2015). Thanks @Uber for the free ride tonight #womeninbusiness #Dublin |
| **Cluster 9: UBER & startups** |
| SaraMorganSF. (April 1, 2015). RT @KiraMNewman: The 10 fastest-growing startups of 2014 - http://t.co/aHcesVIyuJ @Uber @lyft @airbnb @ga @vice |
| Alliotts. (February 6, 2015). Well done @Uber ! Tech #Crunchies 2014: Uber named best overall startup of the year #tech #startup |
| ReeelTV. (April 16, 2015). We are the uberization of television #uber #startup #disruptiveinnovation #innovation http://t.co/RqapbKfaLi |
| **Cluster 10: Support & Help** |
| Sandatucson. (April 9, 2015). @Uber PLEASE A HUMAN BEING CONTACT US for help. YOU GUYS are impossible! WORST support. Impossible to get in via computer. DISASTER |
| Johnbuzzroll. (March 29, 2015). RT @daynaaab: @Uber my boyfriend just got charged for canceling a ride when the driver asked him too... can you help? |
| dickturp1n. (March 28, 2015).,RT @eapbradford: @Uber @Uber_LDN pls help. You need a contact number. #disappointing |

Table A.6: Sample Tweets

## Highlights

- Collaborated with a professional editor and carefully edited the manuscript to address language and grammatical issues as prompted by the editor and reviewers.
- Restructured the .bib file of our submission, correcting inconsistencies in the references section as prompted by the editor.