ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ
SCHOOL OF BUSINESS

ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ
DEPARTMENT OF MANAGEMENT SCIENCE & TECHNOLOGY

BRAND EQUITY ASSESSMENT:

A Computational Model for Mining Consumer Perceptions in Social Media

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in Management Science & Technology

DEMITRIOS E. POURNARAKIS

2017

[Page intentionally left blank]

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ
SCHOOL OF BUSINESS

ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ
DEPARTMENT OF MANAGEMENT SCIENCE & TECHNOLOGY

ΕΚΤΙΜΗΣΗ ΚΑΤΑΝΑΛΩΤΙΚΗΣ ΑΞΙΑΣ ΕΤΑΙΡΙΚΗΣ ΕΠΩΝΥΜΙΑΣ:

Υπολογιστικό μοντέλο εξόρυξης αντιλήψεων καταναλωτών από τα κοινωνικά δίκτυα

Η Διατριβή Κατατέθηκε για την
Απονομή του Διδακτορικού Τίτλου
στην Διοικητική Επιστήμη & Τεχνολογία

ΔΗΜΗΤΡΙΟΣ Ε. ΠΟΥΡΝΑΡΑΚΗΣ

2017

[Page intentionally left blank]

Abstract

The proliferation of Big Data & Analytics in recent years has compelled marketing practitioners and business decision-makers to search for new methods for generating insights when faced to measure brand performance during campaign appraisals. Marketers are constantly asked by upper management to justify the impact of online marketing activities in terms of brand performance in the marketing mix. Financial measures such as sales and profit provide partial indicators of brand performance, leading marketers to also turn towards assessment of intangible market based assets such as brand equity. To address this gap, marketers over the past 20 years have introduced various approaches that assess key conceptual dimensions of brand equity from a consumer perspective.

Prevailing methods focus on the multidimensionality of the construct using confirmatory factor analysis with structural equations modeling for evaluating their models. Although these measurement approaches have and continue to be used for assessment of brand equity during marketing campaign cycles, a well-observed challenge is that these methods heavily rely on traditional data collection and analysis methods, such as questionnaires, face to face or telephone interviews, and have a significant time lag.

Recent years have witnessed a fundamental shift on 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 to 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. New computational tools that help organize, search, and understand this vast amount of information should now be used during decision making.

In line with this paradigm shift, this study introduces a novel approach of CBBE assessment, through the application of big data and machine learning techniques in data collected from social media. We develop and introduce a conceptual model of steps that lays out a proposed method of how fundamental dimensions of CBBE, such as brand awareness and brand meaning, could be optioned from consumer perceptions, starting from a marketing perspective and describe the technical steps necessary to construct the measure. At the core of the proposed model lie topic and sentiment detection approaches to elicit the influential subjects that govern the generation process of CBBE. Instead of estimating brand equity based on low-level online social verbatim features (e.g. frequency of words with negative or positive sentiment polarity) we propose a high-level abstraction computational model which estimates CBBE based on the central discussion topics on online social networks.

The model utilizes a novel genetic algorithm to address the problem of topic clustering in text data. The proposed topic clustering method is anchored on the Latent Dirichlet Allocation (LDA) probabilistic topic modeling framework, aiming at identifying cluster formations that are optimal in terms of semantic coherence. This work focuses on reformulating the clustering problem as a discrete optimization problem within the n-dimensional standard simplex since all the LDA-based data patterns correspond to n-valued probability distribution vectors. In this way, the NP - hard cluster assignment problem reduces to the problem of locating the optimal cluster-centroid positions within the n-dimensional standard simplex given that the number of clusters to be identified is equal to the number of topics to be extracted by the LDA probabilistic modeling technique. The novelty of the proposed genetic algorithm approach lies primarily upon the adaptation of the centroid-based encoding scheme, in the sense that cluster assignments are implicitly extracted by assigning each data point to the nearest cluster center.

To illustrate the validity of our model we employ two case studies, in different business verticals, that collect brand related data from social media channels, such as Twitter, unveiling key CBBE dimensions during marketing performance assessments. In particular the first case study evaluates the model in the mobile telecommunications business sector, focusing on the two leading carriers in the USA, AT&T and Verizon while the second case study applies the model on the mobile app market, unveiling constitutional elements of CBBE relating to the UBER transportation network.

The theoretical and managerial contributions of this research outline the need for marketing researchers and practitioners to adopt state of the art machine learning techniques and methods, when faced to assess specific marketing constructs, such as customer based brand equity. Our research contributes towards this end by outlining a step by step computational model that starts from a marketing perspective, describes the necessary technological steps and concludes with the insights that can be generated.

Επιτελική Σύνοψη

Οι ραγδαίες τεχνολογικές εξελίξεις των τελευταίων χρόνων καθώς και το πλήθος των δεδομένων τα οποία δημιουργούνται με ρυθμό ταχύτερο από οποιαδήποτε άλλη φορά στην ιστορία, έχουν καταστήσει αναγκαία την προσαρμογή των επιχειρήσεων σε αυτό που σήμερα είναι ευρύτερα γνωστό στην Ελλάδα ως «στρατηγική ψηφιακής σύγκλησης». Ο κλάδος ο οποίος καλείται να προσαρμοστεί πιο απότομα απ' όλους, λόγω της ευθείας σχέσης του με τον καταναλωτή, είναι αυτός του Μάρκετινγκ. Το πλήθος των δεδομένων το οποίο είναι διαθέσιμο σε ηλεκτρονικά μέσα, όπως τα κοινωνικά δίκτυα, οι ιστοσελίδες και τα ιστολόγια, δημιουργεί ένα καινούριο κανάλι γνώσης το οποίο πλέον δεν πρέπει να αμελεί κανείς όταν έρχεται αντιμέτωπος με ανάλυση βασικών εννοιών του μίγματος μάρκετινγκ.

Ένα από τα βασικά στοιχεία το οποίο παραδοσιακά έχει πρωταρχικό ρόλο κατά την χάραξη της στρατηγικής μάρκετινγκ είναι η επίδοση που έχουν οι προωθητικές ενέργειες στο καταναλωτικό κοινό. Οικονομικοί δείκτες, όπως οι πωλήσεις και τα κέρδη, προσφέρουν μια μονοδιάστατη εικόνα της αξίας της επωνυμίας, με αποτέλεσμα ο εμπορικός κόσμος να στραφεί επίσης και στην διάσταση του υπολογισμού ενός άυλου αλλά πολύ σημαντικού περιουσιακού στοιχείου, ευρύτερα γνωστό στην βιβλιογραφία ως «καταναλωτική αξία εταιρικής επωνυμίας». Για την αντιμετώπιση αυτού του κενού έχουν εισαχθεί διάφορες προσεγγίσεις τα τελευταία 20 χρόνια, που αξιολογούν βασικές διαστάσεις της αξίας εταιρικής επωνυμίας από τη σκοπιά του καταναλωτή.

Οι βασικές μέθοδοι επικεντρώνονται στον πολυδιάστατο χαρακτήρα της έννοιας, χρησιμοποιώντας παραδοσιακές μεθόδους ανάλυσης για την αξιολόγηση των μοντέλων τους. Οι προσεγγίσεις αυτές αν και εξακολουθούν να χρησιμοποιούνται μέχρι και σήμερα, βασίζονται σε μεγάλο βαθμό σε παραδοσιακές μεθόδους συλλογής και ανάλυσης δεδομένων, όπως ερωτηματολόγια, συνεντεύξεις πρόσωπο με

πρόσωπο ή τηλεφωνικές συνεντεύξεις, και παρουσιάζουν μια πολύ σημαντική χρονική υστέρηση.

Τα τελευταία χρόνια έχει παρατηρηθεί μια θεμελιώδη αλλαγή σχετικά με το πώς οι καταναλωτές επιλέγουν να μεταφέρουν την εμπειρία τους κατά την αλληλεπίδρασή τους με ένα εμπορικό σήμα, υπηρεσία ή προϊόν. Η άνοδος των πλατφορμών κοινωνικής δικτύωσης έχει δώσει σε καταναλωτές σε όλο τον κόσμο την δυνατότητα να επηρεάζουν και να διαμορφώνουν αντιλήψεις σχετικά με την αξία των εμπορικών σημάτων επιχειρήσεων. Το γεγονός αυτό δεν αφήνει καμία επιλογή στους οργανισμούς πάρα να προσαρμοστούν, να επενδύσουν και να παρακολουθούν αυτά τα κανάλια.

Με την ανάπτυξη των φορητών συσκευών, οι οποίες προσφέρουν πρόσβαση στο διαδίκτυο, ανεξαρτήτως χρόνου και τόπου, οδηγούμαστε στην εποχή του συνδεδεμένου καταναλωτή, ο οποίος με τη δύναμη ενός «έξυπνου κινητού» έχει πλέον πρόσβαση σε πληροφορίες που προηγουμένως ήταν διαθέσιμες μόνο μέσω συμβατικών μέσων. Ζούμε την εποχή όπου το μάρκετινγκ και η τεχνολογία είναι περισσότερο από ποτέ κοντά στην σύγκλιση και τελικά την συγχώνευση. Οι πληροφορίες πλέον ψηφιοποιούνται και αποθηκεύονται σε πραγματικό χρόνο στο διαδίκτυο, καθιστώντας ολοένα και πιο σημαντικό για τις επιχειρήσεις να μην παραμελούν τέτοια μέσα. Νέα εργαλεία και υπολογιστικά μοντέλα που βοηθούν στην αναζήτηση, στην οργάνωση αλλά και στην κατανόηση της θάλασσας των δεδομένων είναι πλέον αναγκαίο να χρησιμοποιούνται κατά τη διάρκεια της διαδικασίας λήψης αποφάσεων.

Το κύριο αντικείμενο αυτής της διδακτορικής διατριβής είναι να εισάγει μια νέα μέθοδο για την εκτίμηση καταναλωτικής άξιας εταιρικής επωνυμίας, παρουσιάζοντας ένα υπολογιστικό μοντέλο εξόρυξης αντιλήψεων καταναλωτών από τα κοινωνικά δίκτυα. Για να το πετύχει αυτό χρησιμοποιεί υπολογιστικές μεθόδους τεχνητής νοημοσύνης και ανάλυσης δεδομένων. Πιο συγκεκριμένα εισαγάγει ένα εννοιολογικό μοντέλο διακριτών βημάτων που καθορίζει μια προτεινόμενη μέθοδο, για το πώς θεμελιώδεις διαστάσεις της καταναλωτικής άξιας εταιρικής επωνυμίας, όπως η

αναγνωρισιμότητα του σήματος και η έννοια της μάρκας, θα μπορούσαν να αντληθούν από τις αντιλήψεις των καταναλωτών, ξεκινώντας από την σκοπιά του μάρκετινγκ και περιγράφοντας τα τεχνικά μέτρα που είναι αναγκαία για την κατασκευή του μέτρου αυτού. Στον πυρήνα του προτεινόμενου μοντέλου, βρίσκονται μέθοδοι ανίχνευσης συναισθήματος και προσεγγίσεις απόσπασης βασικών θεμάτων που διέπουν τη διαδικασία δημιουργίας της εκτίμησης καταναλωτικής άξιας εταιρικής επωνυμίας. Το μοντέλο αντί να βασίζεται στον υπολογισμό του μέτρου με βάση ανάλυση δεδομένων από παραδοσιακά μέσα (ερωτηματολόγια), προτείνει τον υπολογισμό βάση των κεντρικών θεμάτων συζήτησης στα κοινωνικά μέσα δικτύωσης.

Πιο συγκεκριμένα το μοντέλο χρησιμοποιεί έναν γενετικό αλγόριθμο για την αντιμετώπιση του προβλήματος της ομαδοποίησης θεμάτων σε δεδομένα κειμένου. Η προτεινόμενη μέθοδος ομαδοποίησης δεδομένων σε συναφή θέματα βασίζεται στην τεχνική " Latent Dirichlet Allocation (LDA)", με σκοπό τον εντοπισμό σχηματισμών συμπλεγμάτων με βέλτιστη σημασιολογική συνοχή. Παράλληλα χρησιμοποιεί μια σειρά από τεχνικές τεχνητής νοημοσύνης για τον εντοπισμό συναισθήματος (θετικού ή αρνητικού) στους σχηματισμούς που ορίζονται από τον γενετικό αλγόριθμο. Με τον συνδυασμό των τεχνικών αυτών, το μοντέλο καταλήγει σε μια σειρά μετρικών οι οποίες αναδεικνύουν δύο σημαντικές διαστάσεις καταναλωτικής άξιας εταιρικής επωνυμίας, την αναγνωρισιμότητα της επωνυμίας και την σημασιολογική βαρύτητα που δίνει ο καταναλωτής.

Για να αναδείξουμε την ισχύ και την ορθότητα του μοντέλου, τρέχουμε δύο μελέτες περίπτωσης που εφαρμόζουν το μοντέλο σε επιχειρήσεις διαφορετικών εργασιών. Στόχος και στις δύο περιπτώσεις είναι μέσα από την συλλογή δεδομένων από το κοινωνικό μέσο δικτύωσης, Twitter, να αναδειχθούν οι διαστάσεις της καταναλωτικής άξιας εταιρικής επωνυμίας μέσα από την διαδικασία ορισμού του αρχικού προβλήματος από σκοπιάς μάρκετινγκ. Πιο συγκεκριμένα, η πρώτη μελέτη αξιολογεί το μοντέλο στον τομέα των επιχειρήσεων κινητής τηλεφωνίας, με επίκεντρο τις δύο κορυφαίες φορείς στις ΗΠΑ, AT&T και Verizon, ενώ η δεύτερη

μελέτη εφαρμόζει το μοντέλο στην αγορά εφαρμογών για κινητά, αναδεικνύοντας διαστάσεις της καταναλωτικής άξιας του δικτύου μεταφοράς UBER.

Η συνεισφορά της παρούσας διατριβής στην βιβλιογραφία είναι πολλαπλή. Η σημαντικότερη συμβολή όμως αυτής της έρευνας έγκειται στην ανάδειξη της ανάγκης υιοθέτησης τεχνικών μηχανικής μάθησης σε προβλήματα στον χώρο του μάρκετινγκ, και την χρησιμοποίησης ενός τέτοιου μοντέλου, όταν οργανισμοί έρχονται αντιμέτωποι με την αξιολόγηση στρατηγικών σχετικά με την άξια εταιρικής επωνυμίας από την σκοπιά του καταναλωτή. Η έρευνά μας συμβάλλει προς αυτή την κατεύθυνση, εισάγοντας ένα τέτοιο υπολογιστικό μοντέλο, που ξεκινά από την πλευρά της αγοράς, περιγράφει τα απαραίτητα τεχνολογικά βήματα και καταλήγει με τις ιδέες που μπορούν να δημιουργηθούν.

Publications

The following papers have been published as a direct or indirect result of the research discussed in this dissertation:

Papers in refereed Journals:

1. Pournarakis, D; Sotiropoulos, D; Giaglis, G. (2016) **"A computational model for mining consumer perceptions in social media"**, *Decision Support Systems*, Volume 93, January 2017, Pages 98-110
2. Sotiropoulos, D; Pournarakis, D; Giaglis, G, (2016) **"SVM-Based Sentiment Classification: A Comparative Study against State-of-the-Art Classifiers"**, *International Journal of Computational Intelligence Studies* (Accepted - forthcoming)
3. Giaglis, G, Bilanakos, C; Georgoula, I; Pournarakis, D; Sotiropoulos, D; (2016) **"Economic, Technological and Behavioral Factors, Affecting the Price of Bitcoin"**, *Ledger* (under review)

Papers in peer-reviewed Conference Proceedings:

1. Sotiropoulos, D; Pournarakis, D; Giaglis, G, (2016) "**A Genetic Algorithm Approach for Topic Clustering: A Centroid-Based Encoding Scheme"**, *Proceedings of the 7th International Conference on Information Intelligence, Systems and Applications, 13-15 July 2016, Halkidiki, Greece*
2. Georgoula, I; Pournarakis, D; Sotiropoulos, D; Bilanakos, C; Giaglis, G;(2015) **Using Time-Series and Sentiment Analysis to detect the Determinants of Bitcoin Prices,** Proceedings of the *9th Mediterranean Conference on Information Systems, 3-5 October 2015, Samos, Greece*
3. Sotiropoulos, D; Pournarakis, D; Giaglis, G, (2015) "**Semantically aware time evolution tracking of communities in co-authorship networks."** *Proceedings of the 19th Panhellenic Conference on Informatics, 1-3 October 2015, Athens, Greece*
4. Sotiropoulos, D; Pournarakis, D; Giaglis, G, (2015) **"Tracking the Evolution of Communities in Co-Authorship Networks: A Semantically Aware Approach"** *Proceedings of the 6th International Conference on Information Intelligence, Systems and Applications, 6-8 July 2015, Corfu, Greece*

5. Bouros, N; Sotiropoulos, D; Pournarakis, D; Giaglis, G, (2014) "**Social Network Analysis Within The ICMB Community: Co-Authorship Networks**" *Proceedings of the ICMB 2014: 13th International Conference on Mobile Business, 2014, 4-5 June 2014, London, UK*
6. Pournarakis, D., Kounavis, C., Sotiropoulos, D., Giaglis, G. (2013) "**AT&T VS VERIZON: Mining Twitter for customer satisfaction towards North American Mobile Operators**." *In the Proceedings of the 12th International Conference on Mobile Business (ICMB 2013), June 10-13, 2013, Berlin, Germany* [1]

[1] Best Paper Award Nominee

Acknowledgments

Embarking on a journey towards a PhD thesis is primarily a lonely quest, a quest of continuous wander and individual endeavor. This journey wouldn't have been possible though, without the help and support of many people, to whom I am profoundly grateful.

First of all, I would like to thank my supervisor, Professor George M. Giaglis, for being the inspiration behind my decision to enter the academic world. George always managed to find time for me, despite his overwhelming duties both as a teacher and vice rector. His advice, encouragement and support have been invaluable throughout the process. For this, I thank him deeply.

I am also especially indebted and grateful to Dr. Dionisis Sotiropoulos for being my research-partner in crime for the past four years. This thesis has been greatly influenced and structured based on his discerning comments and I wish to thank him for bearing with me during the endless hours spent in the lab.

I would also like to express my gratitude to the members of my committee, Professor George Siomkos and Assistant Professor Adam Vrechopoulos. Your assistance and guidance was of outmost importance, especially during the first months of inquiry.

Special thanks go out to my research buddies at Sociomine and fellow colleagues of the doctoral program at Athens University of Economics and Business, who have been a friend in need when things got tough.

Finally a great amount of gratitude goes to the anonymous reviewers from conferences and journals. I don't know who you are, but looking back, your comments and guidance helped me shape my research and findings in the most profound way!

Funded Research

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: Sociomine

Η παρούσα έρευνα έχει συγχρηματοδοτηθεί από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο - ΕΚΤ) και από εθνικούς πόρους μέσω του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» του Εθνικού Στρατηγικού Πλαισίου Αναφοράς (ΕΣΠΑ) – Ερευνητικό Χρηματοδοτούμενο Έργο: Sociomine

Table of Contents

List of Figures

List of Tables

[Page intentionally left blank]

Chapter 1

Introduction

I met Catherine, Chief Marketing Officer in one of the biggest airlines in Europe, in early 2012 during a business meeting. At the time I was working for IBM and had the chance to regularly meet with high level marketing executives from various industries, promoting what was at the time an early prototype of IBM's Watson Social Media Analytics Software. Catherine told me that during the previous quarter her company was voted Best Regional Airline in Europe and while this was greatly received by the board of directors, it also created a huge problem for her and the department.

Indeed in 2011 the company received the *"Best Regional Airline in Europe"* award, at the World Airline Awards, as a result of an online vote by over 18.8 million passengers of 100 different nationalities who selected the best airline out of a total of 200 companies, based on customer satisfaction, as regards 38 different items of front-line product and services[2].

Amazed by her response, I asked how such a fortunate event could be a problem for their department. The answer was honestly shocking.

[2] Names used for the purposes of this thesis are fictional to protect anonymity and confidentiality agreements between the researcher and the parties.

"If I am going to be completely honest with you, we found out we had won this award only after the media picked this up from the web, started publishing it on print media, and contacted us for our comment."

Apparently the airline had not participated in any sort of submission for such a competition and consequently wasn't aware of its nomination and win, which came solely from consumer votes through social media.

Catherine's story was no shock, as similar –but not so blunt- responses were also coming from executives from different market verticals. What did they all have in common? All executives replied that their departments where not ready to handle all this vast amount of information coming from social media, they had no idea how their company was performing on these channels, and their only interaction with these media was through a team of 1 or 2 people who manually browsed Facebook and Twitter (at the time), on a daily basis, to fetch stories relevant to their brand.

At the time this research set off, social media where undergoing through major transformation. Sites like Facebook and Twitter initially launched as social networking sites, where an individual could meet and connect with friends, exchange messages and post status updates or pictures. It was at around that time when brands understood the opportunity to address a significant number of potential consumers through this medium. The response from social networking sites was almost immediate, as with minor tweaks to their business model they offered brands the ability to create company pages, providing them a space where they could directly interact with consumers and target them through display advertisements. This change opened up a whole new chapter in marketing, as now people, through the power of their mobile phone or personal computer, could instantly post their experience with the brand, even at the exact moment of their interaction.

Coming back to Catherine's story, it was evident that companies more than ever, needed to quickly adapt to this emerging shift in consumer – brand interaction and

find ways to monitor brand activity and infer insights from what their consumers where posting on these channels. Seeking to provide an answer to the aforementioned problem, the present research lies at the intersection of two research areas, that of Marketing and that of Machine Learning and Big Data Techniques.

The main theme of this research posits the need to introduce a method for assessing customer based brand equity, shifting from traditional data collection and analysis methods, such as questionnaires, face to face interviews or data collected from online databases. We contend that traditional methods of data gathering provide a static and sometimes skewed measure, whereas assessing brand equity through mining consumer perceptions from social media provides a dynamic and near real time measure of what customers are expressing at the moment of brand interaction.

To set the theme of the rest of this thesis, this chapter acts as an introduction to the presented research and in the next subsections introduces the reader to key concepts and definitions, describes the motivation and rationale behind the approach, sets and grounds the research objectives and presents the methodology followed.

1.1 Research Background

Branding is a vital and essential process for all commercial and non-commercial organizations. From an academic perspective a brand is officially defined as:

> *"the name, term, design, symbol, or any other feature that identifies one seller's product distinct from those of other sellers"*[3]

[3] American Marketing Association Dictionary, [Online] Available: https://www.ama.org/resources/Pages/Dictionary.aspx [Accessed: 03-Aug-2016].

From a business perspective we could argue that a brand essentially encompasses a constellation of qualities that help overcome the unreliability of individuals and enable consumers to better understand the value of a product.

People have the tendency to associate branding with tangible goods such as deodorants, toothpastes, refreshments, cereals, etc. This is partially true as in packaged goods the product is the primary brand the consumer knows and interacts with. Their logos, colors and texture, shape the image the consumers see in the product. However with services, branding is much more different as it essentially is depicted in the image of the company.

The importance of service branding has been thoroughly studied by researchers during the past years (Beverland et al. 2007; D. F. Davis, Golicic, and Marquardt 2008; De Chernatony, McDonald, and McDonald 1992; Lovelock, Wirtz, and Chew 2009; Onkvisit and Shaw 1989). Most of these studies draw on Berry's (2000) conclusion that:

> *"Strong service brands increase customers' trust of the invisible purchase as they enable them to better visualize and understand intangible products. What strong service brands achieve is to reduce customers' perceived monetary, social or safety risk in buying services, which are difficult to evaluate prior to purchase."*

A related and complementary concept that sparked conversation near the end of the 1980's is that of Brand Equity. The nature of the concept is so complex that various conceptualizations of the given construct have been presented and are still published in literature. Despite many attempts to universally define the term, there is little consensus up to today on the exact meaning of brand equity and its aspects. Perhaps the most widely accepted definition is given by Farquhar (1989) as *"the added value that a brand endows a product with".*

Two different schools of thought govern the study of brand equity and its dimensions. One is financial based, which estimates the value of a brand primarily for accounting purposes such as market valuation and goodwill. The other focuses on the strategic aspect of branding in terms of improving marketing productivity by gaining knowledge that has been created about the brand in consumer's minds from the firm's investment in previous marketing actions. This aspect has been coined by Keller (1993) as Customer-based brand equity which is essentially defined as:

> *"the differential effect of brand knowledge on consumer response to the marketing of the brand".*

Table 1.1 summarizes other prevailing definitions of Customer Based Brand Equity as presented by various researchers in literature.

Author(s)	Construct Definition
The Marketing Science Institute (Leuthesser 1988)	The set of associations and behaviors on the part of the brand's consumers, channel members, and parent corporation that permits the brand to earn greater volume or greater margins than it would without the brand name and that gives the brand a strong, sustainable, and differentiated advantage over competitors.
Aaker (1992)	The value consumers associate with a brand, as reflected in the dimensions of brand awareness, brand associations, perceived quality ,brand loyalty and other proprietary brand asset.
Keller (1993)	Customer-based brand equity is defined as the differential effect of brand knowledge on consumer response to the marketing of the brand. A brand is said to have positive (negative) customer-based brand equity when consumers react more (less) favorably to an element of the marketing mix for the brand than they do to the same marketing mix element when it is attributed to a fictitiously named or unnamed version of the product or service.
Moore (1993)	Customer-based brand equity is defined as the combination of brand awareness, liking and perceptions
Park & Srinivasan	Customer-based brand equity may be defined as the added value endowed by the brand to the product as perceived by a

(1994)	consumer.
Lassar (1995)	The enhancement in the perceived utility and desirability a brand name confers on a product based on five dimensions. Performance, value, social image, trustworthiness and commitment.
Berry (2000)	It is a blend of what the company says the brand is, what others say, and how the company performs the service--all from the customer's point of view.

Table 1.1: Definitions of Customer Based Brand Equity in Literature

A direct consequence of this emerging school of thought is the attempt to provide methodologies for measuring customer based brand equity. Two basic approaches have been introduced in literature as shown in figure 1.1.

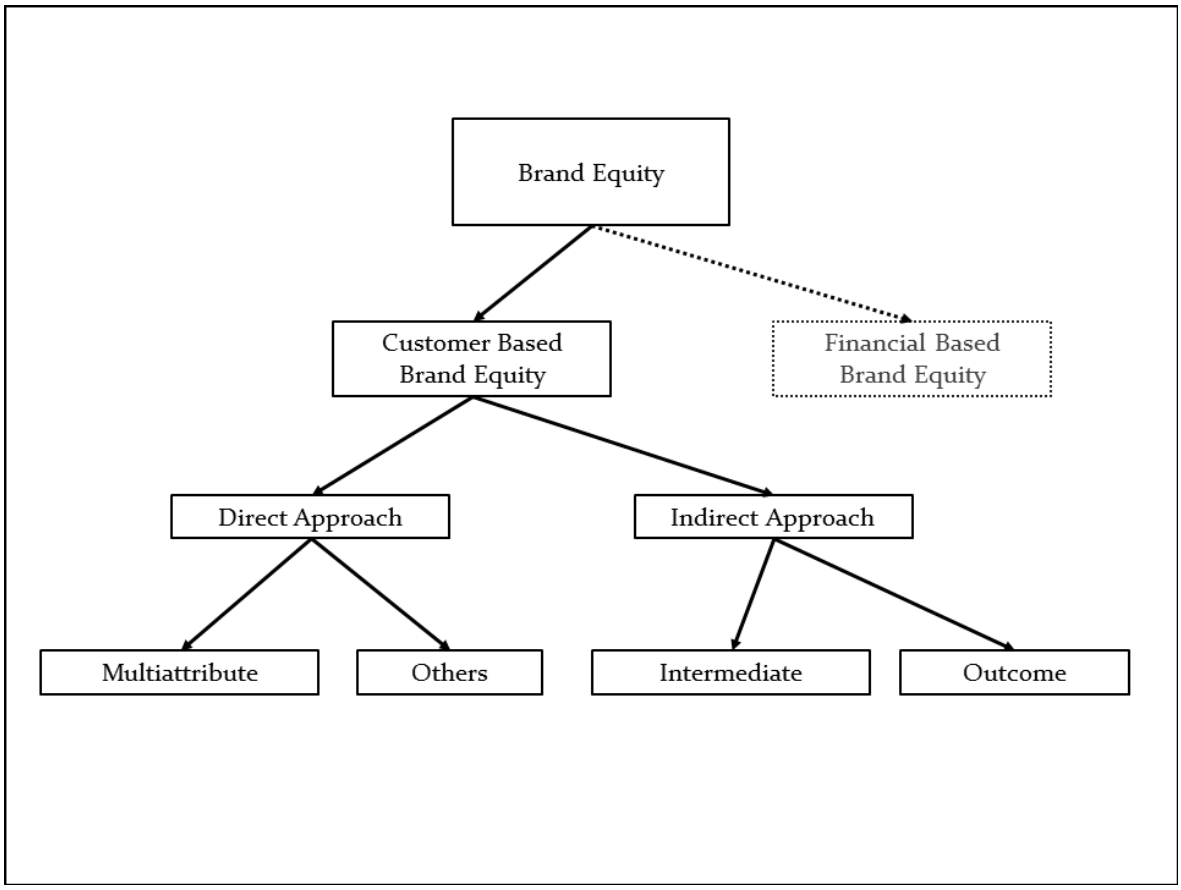


Figure 1.1: Brand Equity Methodologies, (Christodoulides and De Chernatony 2010)

The direct approach, which attempts to measure CBBE more directly by assessing the impact of brand knowledge on consumer response to different elements of the firm's marketing program and the indirect approach, which attempt to assess potential sources of CBBE by measuring its distinct dimensions. Both approaches, which are still used today by marketers, focus on the multidimensionality of the construct, collect data through traditional methods, such as surveys and questionnaires and use confirmatory factor analysis with structural equations modeling for evaluating their models.

During the same period, researchers from the Information Systems discipline raised awareness on the massive proliferation of social media and how this new medium was changing the way consumers interacted with businesses worldwide. This effect sparked continuous interest in regards to the massive amount of data that was generated on a 24/7 basis through these mediums and which companies at the time had no way of analyzing or infer any sort of insight.

In 2009, a study by Melville, Sindhwani & Lawrence was among the first to highlight the need to introduce machine learning techniques to tackle problems relating to automated analysis of data from blogs for marketing purposes. In their study they acutely state:

> *"This rise of the blogosphere has empowered the average consumer with the ability to influence the public perception and profitability of brands. As such, marketing organizations need to be mindful of what people in general (and potential customers in particular) are saying in blogs, how the expressed opinions could impact their business, and how to extract (and drive) business insight and value from these blogs."*

The specific paper is the motivation behind the research presented in this thesis, which is described in more detail in the section that follows.

1.2 Research Motivation & Objectives

Motivation for this paper arose out of the need to introduce a new method for eliciting influential subjects that govern brand equity assessment for service brands, by mining and analyzing consumer perceptions from online social network data. The main objective of this research is to design and evaluate a computational model that lays out a proposed method on how consumer perceptions could be optioned, starting from a marketing perspective, 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.

Our proposed model draws on Aaker's (1992) definition on Brand Equity, utilizes Berry's (2000) conception of CBBE for service brands, using the dimensions of brand awareness and brand meaning to assess CBBE and draws on Bruhn's et al (2012) claim that social media are starting to replace traditional media in terms of brand equity creation.

The study aims to contribute to the literature in the following ways. Our primary aim is to stress the need to complement traditional methods of brand equity assessment with methods that draw upon computational models. Technological advances in machine learning and data analysis should lead in taking the unprecedented step towards a combined marketing-driven approach that utilizes big data and machine learning techniques. Through the study, we aim to stress the need for marketing and information systems (both as scientific disciplines and within the organizational context) to be intertwined in order to compute, assess and interpret consumer perceptions towards brands and reveal insights regarding customer based brand equity dimensions. To do so we utilize and aim to improve current data clustering methods to extract semantically focused groups of documents when compared against traditional clustering algorithms such as the k-means.

From a practitioner's perspective, our objective is to provide managers with an actionable method that can be utilized and applied in daily operations. 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. 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.

1.3 Research Methodology

The research inquiry governing the thesis may be classified as of exploratory nature as it aims to aid in identifying a specific problem, clarify the nature of it and define its scope (McGivern 2009). It could also be viewed as adhering to the interpretive paradigm of social science research, employing an inductive approach that starts with data and tries to derive a theory about the phenomenon of interest from the observed data (Bhattacherjee 2012). In particular the thesis explores how raw streams of data elicited from Social Media Networks (SMN) can be used to derive quantifiable customer based brand equity insights during assessment of service brand actions.

A combination of different research methods were followed as part of this research. This presented the need for the methods to fall under a potent research framework that would allow for an apt and efficient approach in identifying the problems under investigation. As such the research approach for Information Systems by Galliers (1992) was selected. Figure 1.2 illustrates the approach followed in this thesis compared to Gallier's generic IS approach.

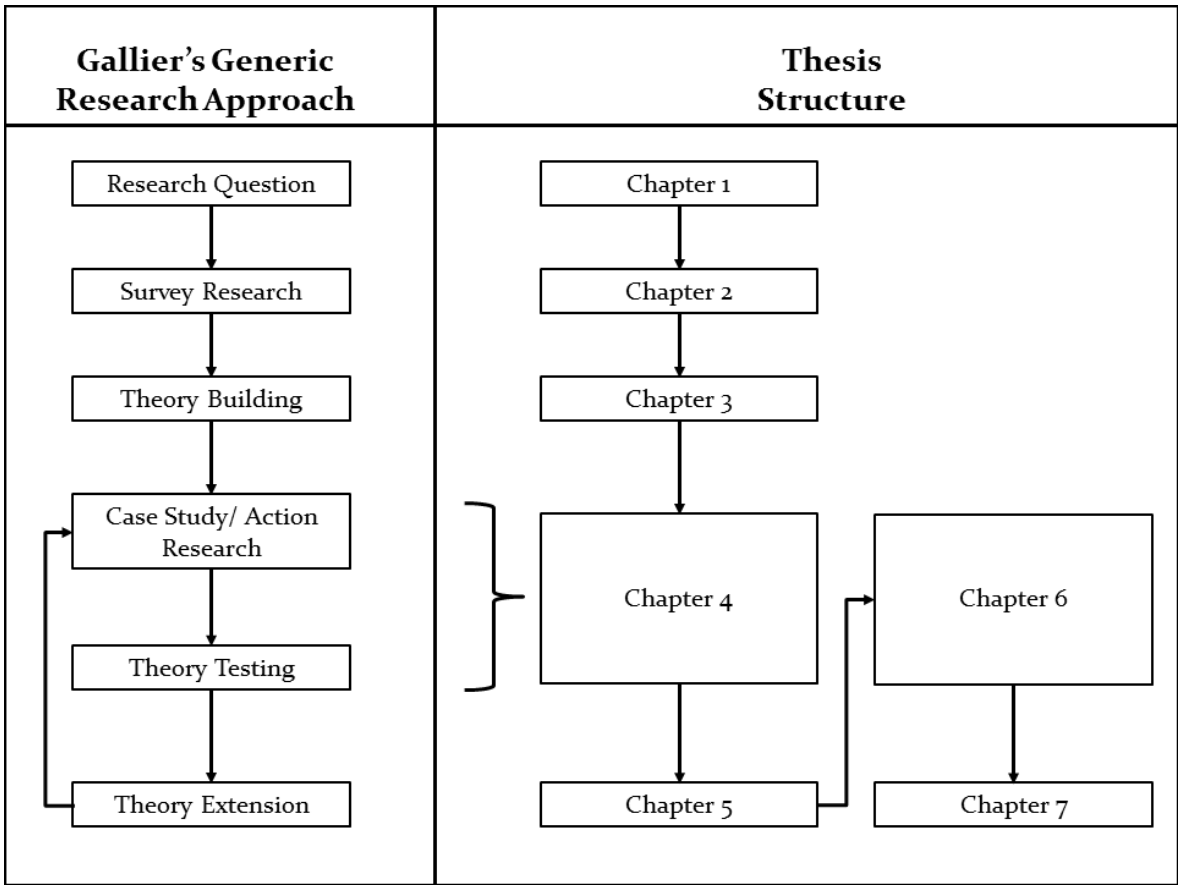


Figure 1.2: Gallier's Generic IS Research Approach(1992) vs. Thesis Structure

The results of the first case study highlighted the need to refine the model, and ground the conceptual framework for understanding, executing, and evaluating IS research through the design-science paradigm as proposed by Hevner et al (2004), in line with guidelines introduced (chapter 5). The overall goal was to develop and introduce a conceptual model of steps that lays out a proposed method of how customer based brand equity assessment could be optioned via Social Media Analytics, starting from a business perspective, then describing the technical steps necessary to construct the measure, concluding with how the results could be interpreted. The utility, quality, and efficacy of the design artifact were evaluated through conduct of a second case study, assessing the artifact in depth across a different market vertical in the business environment, seeking to achieve external validity and generalizability of the method (Bhattacherjee 2012) as presented in figure

1.3. Finally, planning and timescale of the research process followed the framework proposed by Philips and Pugh (2005).

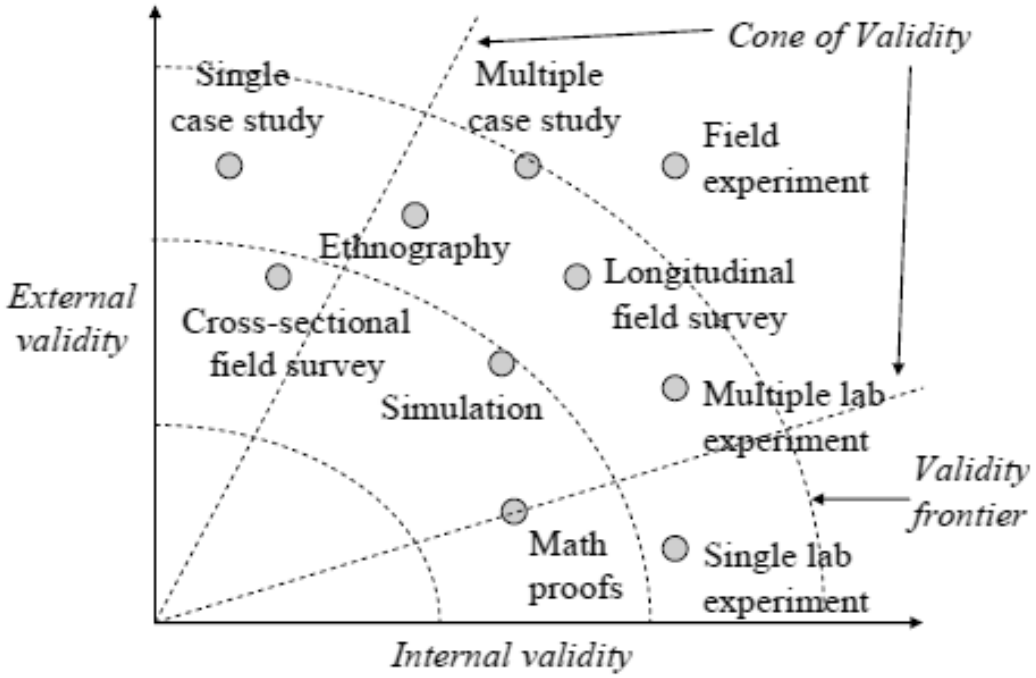


Figure 1.3: Cone of Validity (Bhattacherjee 2012).

1.4 Thesis Outline

The research presented in this thesis is structured around seven distinct chapters.

Chapter 1 introduces the reader to key concepts and definitions, describes the motivation and rationale behind the suggested approach, sets and grounds the research objectives and presents the methodology followed.

Chapter 2 provides an in-depth analysis of the theoretical background that draws upon three scientific disciplines. Through a thorough literature review the chapter presents relevant prescriptive theory and existing artifacts of brand equity and Social Media Analytics techniques, as well as descriptive theory and prior knowledge on these constructs. The chapter continues by drawing upon previous work

published in prominent journals which either assess the problem of measuring customer based brand equity or present methods of applying social media analytics that assess streams of text into topics and sentiment scales. This review derives the need to harness the growing amount of data generated from social media channels from an IT perspective and shift towards measurement techniques that leverage such data.

Chapter 3 introduces the theoretical framework and the development of a primary conceptual model which describes the appropriate process of gathering and analyzing data from online social networks. In particular it assists researchers in mining online social media data-streams in order to detect trends or patterns of consumer behavior and analyze structural relationships. This chapter serves as the necessary technical background to conduct a series of experiments and test the validity of the theoretical predictions in the next steps of this research.

Chapter 4 presents an empirical case study, with primary intent to test the efficiency of the framework introduced in the previous chapter. The chapter documents and explores how mobile wireless carriers can benefit from performing topic modeling and sentiment analysis from social media data in order to enhance and improve customer satisfaction scores. The findings of this case study set the base of discussion and highlight the need for refinement, providing the specifics for enhancement of the theoretical model to be discussed in detail in the next chapter.

Chapter 5 introduces the refined method for eliciting influential subjects that govern brand equity assessment, by mining and analyzing consumer perceptions from online social network data. It does so by grounding the research through a design science research approach. The chapter describes the design and evaluation of a computational model that lays out a proposed method on how consumer perceptions could be optioned, starting from a marketing perspective, the technical steps necessary to construct the assessment metrics, and concludes with how the results could be interpreted and utilized in brand equity assessment exercises.

Chapter 6 employs a real-life case study of one of the leading service brands in the transport and app ecosystem (UBER). This time the case study is structured around the computational model introduced in the previous chapter and follows the suggested approach through a step by step guide. The study concludes by presenting 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. 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.

Chapter 7 concludes with the contributions of the proposed research and discusses comments for future direction and development in the area of digital marketing and brand equity assessment.

1.5 Summary

This chapter aimed at providing an introductory discussion of the research presented in this thesis. In particular it set off to highlight the key definitions, concepts and scientific disciplines that the research draws upon and that will be discussed in the chapters to follow. We also shed light behind the motivation that sparked this journey and the objectives set forth. After a brief presentation of the research methodology the chapter outlined the structure of the thesis that will guide the reader in the remaining parts to follow.

Chapter 2

Review of Literature

Social media platforms and web 2.0 applications have recently been identified as the key drivers behind the fundamental shift in marketing research methods and brand assessment exercises. This chapter reviews current literature related to marketing and machine learning and discusses how these two disciplines intertwine during brand analysis exercises. In particular we briefly set the tone of current research regarding the problem of market based analysis in the introduction section, while section 2.2 reviews current methods followed by marketing researchers during assessment of customer based brand equity. Section 2.3 discusses and reviews literature on the shift in consumer & brand interaction as a result of the rapid rise and use of social media. The chapter concludes with a thorough review of current big data analytics & machine learning techniques that deal with data and sentiment classification analysis, as introduced in computer science literature.

2.1 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. As Keller (1993) acutely states:

"perhaps a firm's most valuable asset for improving marketing productivity is the knowledge that has been created about the brand in consumers' minds from the firm's investment in previous marketing programs."

This concept of ability to experience, identify or understand customers' sentiment towards brands derives from the notion of empathy as introduced by psychologists in the 19th century. Empathy has been identified by some as primarily an affective phenomenon referring to the immediate experience of the emotions of another person. Others, however, view empathy as primarily a cognitive construct referring to the intellectual understanding of another's experience, while a third view holds that empathy contains both cognitive and affective components (Duan and Hill 1996) . Measuring empathy has been a difficult task for many psychologists. Attempts to measure empathy include self-reports reports of patients, observer ratings and physiological measures but most if not all of the methods are problematic because they fail to differentiate empathy from other emotions such as sympathy and personal distress.

In their attempt to conceptualize this notion, researchers from marketing focus attention on the construct of brand equity which refers to the added value consumers place to a product or service (Yoo and Donthu 2001) and has been most comprehensively defined by Aaker, Keller and Berry (1996; 1993; 2000) as part of their studies during the 90's. As marketers constantly seek ways to justify the impact of online and offline marketing activities in terms of brand performance in the marketing mix, especially in terms of consumer adoption and with financial measures providing only partial indicators of brand performance, it still remains a challenge to justify marketing actions as part of an overall business strategy

Several researchers and marketing practitioners have conceptualized brand equity similarly to Aaker and Keller and used the term consumer-based brand equity, introducing measurement methods based on qualitative assessments or surveys. The strategic impact of branding is duly recognized in the marketing literature (Kapferer 2012; McDonald, de Chernatony, and Harris 2001) as researchers have

alluded on how vital, valid and reliable consumer-based brand equity instruments have become for brand managers. Nevertheless there is still no general agreement among marketing researchers, at the conceptual level about what brand equity truly comprises (Pappu, Quester, and Cooksey 2005) and how valid these measures are.

One of the challenges of current practices that try to assess this problem is that they heavily rely on traditional data collection and analysis and thus have a significant time lag (Christodoulides and De Chernatony 2010). Recent years though have witnessed a fundamental shift on 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 to 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 (Blei 2012).

Social media networks (SMN), like Facebook or Twitter, have long been studied as channels through which consumers may convey their experience or sentiment against brands. As identified by prior studies (Aral and Walker 2011; Palka, Pousttchi, and Wiedemann 2009), consumer reactions towards brands can have viral effects in SMN, resulting in potentially disastrous consequences for companies who miss or avoid monitoring social media channels (Dellarocas and Wood 2008; Hoffman and Fodor 2010; Luo, Zhang, and Duan 2013). This proliferation of social media has sparked a fascination with what Surowiecki (2005) called, in a similarly titled book, “The Wisdom of Crowds”. Behind this fascination lies the, admittedly attractive, idea that by somehow aggregating the imperfect, distributed information held by individual members of a large social network we can yield more valid knowledge than that obtained through experts or artificial markets. As such,

organizations are exploiting all available means that could provide insights about customer-based brand equity, recently drawing upon social media analytics (SMA) methods and techniques. By applying SMA in large sets of data originating in social media, organizations can obtain knowledge on organizational objectives, such as brand equity.

Whereas prior studies attempting to measure brand equity rely predominantly on measures based on traditional data collection methods, such as questionnaires, face to face interviews or data collected from online databases (Lassar, Mittal, and Sharma 1995; Park and Srinivasan 1994; Swait et al. 1993), there is the need to introduce more valid and timely results by applying SMA techniques in social media datasets. To this end this study seeks to expose the interrelation between different disciplines, such as Marketing, Information Systems and Computer Science, highlighting relevant effects between them and poses to argue that computing algorithms and machine learning techniques should be utilized to answer marketing problems in terms of brand equity assessment. The interrelation between disciplines highlights the need for organizations to consider models where strategic objectives are measured, analyzed and improved, through cross sectional collaboration between departments.

The rest of the chapter presents relevant prescriptive theory and existing artifacts of brand equity and SMA techniques, as well as descriptive theory and prior knowledge on these constructs. We draw upon previous work published in prominent journals which either assess the problem of measuring customer based brand equity or present methods of applying social media analytics that assess streams of text into sentiment scales. We also draw upon key research papers which highlight the need to harness the growing amount of data generated from social media channels from an IT perspective and shift towards measurement techniques that leverage this data. The remaining part of this section is broken down to three subsections which will further aid to theoretically ground our model.

1. Current & Previous work on measuring Customer Based Brand Equity

2. Proliferation of social media and rapid growth of social and mobile commerce as a driver for fundamental shift between consumer and brand interaction
3. Current & Previous work on Data Clustering and Classification

2.2 Current & Previous Work on Customer Based Brand Equity Assessment

The brand as a concept has been the subject of much research within the marketing community especially in regards to the construct's equity (Aaker, David A. 1992; L. L. Berry 2000; Keller 1993). Despite many attempts to define the term, there is little consensus up to today on the exact meaning of brand equity, nor is there general agreement at the conceptual level about what brand equity comprises and how it may be measured (Pappu, Quester, and Cooksey 2005; Yoo and Donthu 2001).

Perhaps the most widely accepted definition up to today stems from the works of Berry (2000) and Richards (1998) who argue that strong brands act as a safe haven for customers, enable them to better visualize and understand products and promise future satisfaction.

Based on the above, this research adopts the definition as introduced by Berry (2000) and Keller (1993) which states that brand equity may be defined as the differential effect of brand awareness and meaning, combined on customer response to the marketing of the brand. Basically, brand equity stems from the greater confidence that consumers place in a brand than they do in its competitors (Lassar, Mittal, and Sharma 1995)

The need to build high levels of brand equity along with the strategic impact of branding is considered to be essential and has been reported to bring numerous advantages to any organization when running any form of marketing campaign (Kapferer 2012) . For example, Keller and Aaker have both stressed the value that brand equity provides to organizations, namely by aiding in high consumer

preferences, purchase intentions, high stock returns, sustainable competitive advantages, brand loyalty, increased use of satisfaction and enhanced information processing. Keller (1993), classifies these advantages in to two discrete groups, accordant to their motivation. One is financial based, stressing the value of the brand to the firm (Mahajan, Rao, and Srivastava 1990; Simon and Sullivan 1993), while the latter is strategy based, concentrating on the value that has been created about the brand in the consumer's mind (Washburn and Plank 2002; Yoo and Donthu 2001) and coined by Keller as the term also known as Customer-based Brand Equity.

As a consequence, marketers need a more thorough understanding of consumer behavior as a basis for making better strategic decisions about target market definition and product positioning, as well as better tactical decisions about specific marketing mix actions. Perhaps a firm's most valuable asset for improving marketing productivity is the knowledge that has been created about the brand in consumers' minds from the firm's investment in previous marketing programs. As Farquahar (1990) eloquently puts it:

> *"the competitive advantage of firms that have brands with high equity includes the opportunity for successful extensions, resilience against competitors' promotional pressures, and creation of barriers to competitive entry".*

Berry and Keller in their studies, argue that brand equity is measurable and can be positive or negative. Positive brand equity is the degree of marketing advantage a brand would hold over an unnamed or fictitiously named competitor while negative brand equity is the degree of marketing disadvantage linked to a specific brand.

Attempts to provide brand equity measurements based on the consumer perspective have seen light in marketing literature during the past two decades. Prevailing methods focus on the multidimensionality of the construct using confirmatory factor analysis with structural equations modeling for evaluating their models. Cobb-Walgren et al. (1995) were amongst the first to introduce a measurement

approach to consumer-based brand equity based on the conceptualization of Aaker and Keller as a set of four dimensions, namely brand awareness, brand associations, perceived quality and brand loyalty. Yoo and Donthu (2001) treated consumer-based brand equity as a three-dimensional construct, combining brand awareness and brand associations, Washburn and Plank (2002) build on their scale by extending the development, while Pappu et al (2005) suggest again a four-dimension model which incorporates brand personality measures. Similarly a number of marketing researchers (Christodoulides et al. 2006; Tong and Hawley 2009; French and Smith 2013) propose CBBE measurement methods, each assessing different dimensions to achieve the expected outcome. Table 2.1 summarizes studies that are widely used during CBBE assessment along with the dimensions of brand equity that each study draws upon.

Authors	CBBE Dimensions
Aaker (1996)	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
Srinivasan and Park (1994)	Attribute based, Non attribute based
Rego et al. (2009)	Familiarity, Perceived Quality, Purchase

	Consideration, Uniqueness
De Chernatony et al. (2004)	Reputation, Satisfaction, Brand Loyalty

Table 2.1 CBBE Dimensions in Literature

Although these measurement approaches have and continue to be 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 the need to assess and analyze data originating from Social Media Networks (SMN) (Melville, Sindhwani, and Lawrence 2009; Kane et al. 2014; Luo, Zhang, and Duan 2013). Although several frameworks revealing the need to apply SMA techniques have been proposed (Liang and Turban 2011; Fan and Gordon 2014), few manage to provide methodologies for valuating precise marketing constructs that could be assessed by SMA (Hassan Zadeh and Sharda 2014; Callarisa et al. 2012; Luo, Zhang, and Duan 2013; Yu, Duan, and Cao 2013) and mainly focus on the financial aspect of Brand Equity or Firm Equity.

Researchers from Marketing (Coulter et al. 2012; Culotta and Cutler 2016) have recently stressed the need to deliver actionable insights from social media sources. This new information source is being used to better understand customers, improve future operations and provide metrics that can assist CBBE measurement exercises. This allows the brand to gain the advantage of a holistic and near real time view of their customers, complementing current approaches that heavily rely on traditional data collection and analysis methods such as questionnaires, face to face or telephone interviews, which have a significant time lag.

This shift has resulted in a new pool of data that can be mined in order to understand customer perceptions towards the brand. Although positive attributes of customer perceptions can be extracted by various indicators, including current practices, negative attributes can primarily be extracted from consumer perceptions generated from social media due to the nature of the medium that embraces direct

communication. 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.

2.3 Social Media as a Shift in Consumer - Brand Interaction and Data Generation

Social Media, Web 2.0 platforms, smartphones and tablets equipped with LTE and Wi-Fi capabilities have provided consumers with a direct mean to express their experience at the exact time of brand exposure and are fundamentally changing consumer to brand interactions (Gallaugher and Ransbotham 2010; Luo, Zhang, and Duan 2013). Smartphones and Tablets are accelerating this change driving technology and mobile sectors to converge, with social media being a cornerstone in this shift towards an itinerant consumer who in contrast to the past, now has power to shape and influence opinions in regards to brands through a few taps on the keyboard. We live in an age where social media content is updated on a real time basis, with thousands of tweets, posts, shouts, check-ins, etc., being generated every second.

This fundamental change in scale is depicted in figures from various business research studies[4][5] showing the difference in global annual unit sales between conventional desktop PC's with smartphones and tablets (250m units vs. 1.7b units as of 2013). These figures portray a fundamental shift of consumers' needs, questioning the need to use a PC for personal use when users can perform the same activities through their personal smartphone or tablet. Unlike the desktop web,

[4] "Worldwide Mobile Phone 2013–2017 Forecast and Analysis," www.idc.com. [Online]. Available: http://www.idc.com/getdoc.jsp?containerId=239867. [Accessed: 05-May-2014].

[5] "Predicts 2014: Don't Try to Prevent the Digital Revolution, Exploit IT Now." [Online]. Available: https://www.gartner.com/doc/2642621/predicts--dont-try-prevent. [Accessed: 05-May-2014].

smartphones are inherently social. Social Messaging apps have more than a million downloads on Google Play[6] while social media sites like Facebook and Twitter share more than 3 billion pieces of content each day[7] (number of posts growing by day).

This makes consumers of today significantly different to those of 20 or 30 years ago, mainly due to the vast amount of information available, through the tap of a button, in seconds. It is more than ever evident that consumers now create new meanings and values about products and brands in social media that are beyond the control of companies. As Zwass (2010) acutely points out:

> *"internet and social media provide a channel for collective expression of sentiment. What used to be communication between a customer and a firm can now easily be published in social media and attract the attention of thousands of existing or potential customers"*

Companies worldwide seem to be embracing this fundamental shift in commerce. According to an IBM study on over 4000 C-suite executives worldwide, executives consider technology the single most important external force shaping their organizations, while more than half of them say customers now have a considerable influence on their enterprises ("IBM Insights from the IBM Global C-Suite Study" 2014). These enterprises realize that opinions and reviews being shared on social sites is tantamount to customers demanding to be heard and don't hesitate to point out that 82% of the CMO's asked, feel underprepared for the data explosion ("IBM Global CMO Study" 2014).

---

[6] "Mobile is eating the world", Benedict Evans. [Online] Available: http://ben-evans.com/benedictevans/2014/10/28/presentation-mobile-is-eating-the-world. [Accessed: 28-July-2016]

[7] "Social Media Comparison Infographic", Leverage New Media. [Online] Available: https://leveragenewagemedia.com/blog/social-media-infographic/ [Accessed: 28-July-2016]

In a recent report published by McKinsey (Chui et al. 2012), estimations show that the economic impact of social media on business could exceed $1 trillion during the next couple of years. As identified by prior studies (Aral and Walker 2011; Palka, Pousttchi, and Wiedemann 2009), consumer reactions towards brands can have viral effects in SMN, resulting in potentially disastrous consequences for companies who miss or avoid monitoring social media channels (Dellarocas and Wood 2008; Hoffman and Fodor 2010) for purposes such as marketing or knowledge management..

Researchers from the Information Systems(IS) discipline have recently alluded the need to assess and analyze data originating from social media networks, such as Facebook and Twitter (Fan and Gordon 2014; Kane et al. 2014; Melville, Sindhwani, and Lawrence 2009; Zeng et al. 2010). In order to understand the concept of social media networks, we borrow the definition proposed by Kane et al (2014), which defines SMN as "a possession of four essential features", namely:

1. Users have a unique user profile that is constructed by them, by members of their network, and by the platform;
2. Users access digital content through, and protect it from, various search mechanisms provided by the platform;
3. Users can articulate a list of other users with whom they share a relational connection;
4. Users view and traverse their connections and those made by others on the platform.

One of the most studied characteristics and effects of SMN is Word of Mouth (WOM). In their study, Amblee & Bui (2011) argue that WOM communications have been shown to influence customer attitudes and behavior towards a given topic or brand, which in turn can be either positive or negative. Tirunillai and Tellis (2012), complement this argument that consumers seeking information about a potential future buy are very likely to be influenced by their peers, network and comments of previous buyers of the product. Monitoring consumer feedback and WOM through

social media channels and responding accordingly, can act as a supplementary mean of maintaining high customer satisfaction levels. Davidson & Capulski (2006), prove that feedback from reviews can affect brand reputation while Luo et al (2013) investigated this argument to prove that social media metrics may allow investors to not only monitor the firm's customer sentiment and brand performance but also predict its future business value.

Earlier studies have pointed the need for organizations to derive knowledge from SMN. Jevons & Gabbott in their study (2000) were among the first to highlight this opportunity. The significance and importance of being able to monitor what customers are saying and what opinions they express has been highlighted by research performed by IBM. In particular Melville et al (2009) researched customer feedback from various blogs, and laid the ground rules for inferring marketing insights from online data. This argument has quickly escalated the need to expand research not only to blogs but rather the majority of SMN. As such it is evident that as SMN become ubiquitous and information sharing explodes, their content can be mined and the resulting information can be used to infer emerging social behavior and depict a real-world state of human sentiment towards trends.

Early efforts to experiment with data mining in SMN and derive knowledge on business measures include the use of twitter feeds to forecast box-office revenues for movies (Asur and Huberman 2010), election results (Tumasjan et al. 2010), Oscar predictions ("Oscars Senti-Meter" 2014), team and player support during the Super Bowl ("Super Bowl Analysis Takes Us Beyond the Tweets" 2014), brand post popularity (Hassan Zadeh and Sharda 2014) and the examination of blog data to predict spikes in book sales (Gruhl et al. 2005). Similarly, Bollen et al (2011) present a fascinating, yet subject to significant limitations and assumptions, attempt to extract the prevailing social mood from twitter feeds and correlate it to Dow Jones Industrials Average (DJIA) closing values – interestingly, the authors conclude that one social mood dimension (namely calmness) can conditionally be used to improve the predictive accuracy of simple market forecasting models.

A particular field of SMN that has recently enjoyed increasing popularity in organizational research and the computer society is Social Media Analytics (SMA) and more recently, social media intelligence (Zeng et al. 2010). SMA may be defined as the discipline which draws from Social Network Analysis, and combines Machine Learning, Data Mining, Information Retrieval and Natural Language Processing (Melville, Sindhwani, and Lawrence 2009).

However like other socio-technical phenomena, SMA triggers also a dystopian rhetoric (boyd and Crawford 2012). Markets may have already embraced this phenomenon but social scientists still remain critical, especially in regards to ethical, legislation and access limitations.

From a philosophical standpoint, Berry (2011) argues that analyzing data from online sources is merely an exercise of interpreting human behavior and contains *"knowledge and information that lack the regulating force of philosophy"*. This brings up the question to whether analyses that utilize SMA techniques tend to neglect or sideline other methods of consumer behavior and focus only on the quantitative standpoint.

SMA as a method is also prone to claims against its objectivity and accuracy. A large number of researchers debate that results derived from such means, is subjective and what it quantifies is merely a projection of the objective truth. Boyd and Crawford (2012) argue that *"too often, big data enables the practice of apophenia: seeing patterns where none actually exist, simply because enormous quantities of data offer connections that radiate in all directions"*. However the key driver behind this claim is the depth of analysis applied and interpretation of the results, as happens with methods that involve different collection forms of data.

Another factor that should be taken into account is that most SMA analyses are limited to a single platform for collection of data, Twitter in most cases. Twitter is the most popular source for researchers when it comes to data retrieval, but along with its ease of use in obtaining data come a number of methodological challenges.

As Twitter is only part of a wider online platform ecosystem, its user base can at no point be representative of the global population. Nor can all accounts be assumed to be managed by a single person or even a real person (case of bots, aggregators, etc.). Each social media platform is governed by its own set of norms and interactions and may not be representative of general human social behavior (Tufekci 2013). Twitter is widely known and preferred by users for its real-time flow of information, limited message length and its undirected network structure, where users are free to connect with another user without the latter having to connect back. The above do not imply that Twitter is not a suitable network to study. However researchers should be highly skeptical when faced to analyze consumer behavior in forms of status updates, likes, re-tweets, inclusion of geolocation, etc.

Finally a large amount of debate has sprung, regarding the issues involved in the ethics behind data mining and analysis of public content in social media (Buchanan 2012). Even if a user consents to publically broadcast a message, the issue about how ethical it is by anyone to use this information without consent is still debatable by many. The issue of protecting and anonymizing user data in order to protect user's rights is still at a very early stage.

2.4 Data Clustering and Classification

Inferring insights from analyzing vast pools of data has been thoroughly examined in Computer Science literature, and spans throughout a wide range of relevant subfields such as: (a) text normalization, (b) corpus vectorization, (c) sentiment analysis, (d) topic modeling and (e) document clustering.

Text normalization (Clark and Araki 2011; Mikheev 2000) 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 (Salton, Wong, and Yang 1975), 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 (da Silva, Hruschka, and Hruschka 2014; Salehan and Kim 2015; Fersini, Messina, and Pozzi 2014), 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 (Wilson, Wiebe, and Hoffmann 2005; Cai et al. 2008) focuses on identifying the prevailing topics of discussion in a given corpus. Finally, document clustering (Sotiropoulos et al. 2014; W. Xu, Liu, and Gong 2003; Song and Park 2006) addresses the problem of grouping together semantically similar documents which is an imperative prerequisite for estimating the average sentiment value per topic discussed.

Clustering is a fundamental unsupervised classification technique, which has been thoroughly defined in literature (Anil K. Jain 2010; A. K. Jain, Murty, and Flynn 1999; R. Xu and WunschII 2005) as the grouping of a set of k vectors inside a multidimensional space, into n regions of similar characteristics, prior to any knowledge. Its main goal is to classify the number of objects into groups by maximizing their similarity within clusters, while at the same time differentiating them from other distinct groups with respect to a given measure. This technique has sparked interest in a wide range of scientific fields, which turn to clustering techniques, in order to solve complex classification problems. Very interesting cases are proven in neurocomputing (He and Tan 2012), market research (Green, Frank, and Robinson 1967), image segmentation (Scheunders 1997), complex-networks (Pizzuti 2012), pattern recognition (Garai and Chaudhuri 2004) and data mining (Clifton, Cooley, and Rennie 2004; Guha et al. 2003) to name a few.

The techniques used, are broadly categorized in literature as hierarchical and partitional (A. K. Jain, Murty, and Flynn 1999), with the hierarchical method being further classified to agglomerative and divisive (Premalatha and Natarajan 2010; R. Xu and WunschII 2005). In essence, agglomerative methods propose the initial formation of individual distinct clusters, with the two most similar clusters iteratively merging until some termination criterion is met. Divisive methods on the

other hand initially group all vectors in a single cluster while iteratively split into smaller clusters until the objective function is met. The main problem with hierarchical methods is that a wrong choice of merge early on, never leads to an optimal solution. Partitional methods, such as the K-means algorithm, although solve various clustering problems successfully (Aggarwal and Zhai 2012), have serious drawbacks in terms of execution, optimal solution and clustering results (Selim and Ismail 1984; Tou and Gonzalez 1974; Zhong and Ghosh 2005).

Document clustering and topic modeling through either LDA, or probabilistic Latent Semantic Analysis (pLSA) when combined, prove a good fit towards identifying cluster formations that are optimal in terms of semantic coherence. A widely studied approach towards topic clustering is based on matrix factorization. As the name suggests, the basic idea is to transform documents to a latent space, which in turn aid in discovering features underlying the interactions between two different kinds of entities(Aggarwal and Zhai 2012). Prevailing techniques in this domain include Non-Negative Matrix Factorization (NMF) and Latent Semantic Indexing (LSI). In particular, Xu et al. (2003) propose a document clustering method based on NMF, where each axis captures the base topic of a particular document cluster. Similarly Shahnaz et al. (2006) propose a new hybrid technique for NMF, eliminating the need to use subtractive basis vector and encoding calculations. Document clustering with LSI was initially presented by Deerwester et al. (1990) . Song et al. (2009) introduce a GA approach based on LSI, which automatically evolves the proper number of clusters. Kuhn et al. (2007) introduce Semantic Clustering, a technique based on LSI and clustering to group software code artifacts into clusters with similar vocabulary.

Application of Topic Models in document clustering exercises have been introduced in literature, as previously stated, through pLSA and LDA methods (Hofmann 1999). Extensions based on these two methods have been proposed by many researchers (Lu, Tseng, and Yu 2011; Mei and Zhai 2006; Tou and Gonzalez 1974),

such as the LapPLSI method by Cai et al. (2008) and "A segment-based approach to clustering multi-topic documents" by Taggareli & Karypis (2013).

A natural approach in solving problems with high computational complexity, without use of K-means, is through use of evolutionary computing algorithms. Introduced by Holland (1973), Genetic Algorithms (GAs) are a form of stochastic optimization method that derives from Darwin's theory on natural selection. In essence GAs perform adaptive and efficient search in complex and multimodal search spaces and provide near-optimal solutions for objective or fitness function of an optimization problem (L. Davis 1991; Goldberg 1989).

Bandyopadhyay & Pal (2007) provide a rather eloquent description of the essential components and operation of a GA. Every GA is defined by the following set of components according to their definition: A representation strategy that determines the way in which potential solutions will be coded, a population of chromosomes and an accompanied mechanism for evaluating them, a procedure for selection/reproduction via a set of genetic operators and finally the probabilities of performing those operations. GA's operate through a cyclical process, in which each chromosome is selected, evaluated to get a fitness value and genetically manipulated to create a new population of chromosomes. This applies over a number of iterations till one or more termination criteria are satisfied. Either the average fitness value of a population becomes more or less constant over a specified number of generations; either a desired objective function value is attained by at least one string in the population; or the number of iterations is greater than some predefined threshold.

Ever since, GAs have received much attention from the scientific community, as contrary to other clustering algorithms, solve problems through multiple solutions simultaneously. This method has sparked particular interest in the machine learning society (Goldberg and Holland 1988) with scientists proposing use of GAs to solve a wide variety of clustering problems. In particular Blas et al. (2012) propose a grouping evolutionary approach which uses concepts of grouping encoding and

novel adaptations of evolutionary operators, while Chang et al. (2009) propose a new clustering algorithm, based on GA, with gene rearrangement where each chromosome represents the centers of the clusters by a sequence of real-valued numbers. Garai & Chaudhuri (2004) propose a two-phase process GA and Wu & Hsieh (2010) extend the notion by introducing a two-stage approach to story segmentation and topic classification of broadcast news. Tsai et al. (2013) perform feature selection and instance selection based on genetic algorithms using different priorities to examine the classification performances over different domain data-sets. Several works (Murthy and Chowdhury 1996; Song, Li, and Park 2009; Song Wei and Soon Cheol Park 2009; Song and Park 2006) propose the use of GAs to solve clustering problems, by automatically finding a proper value of the number of clusters, while others solve the problem by pre-defining them (Maulik and Bandyopadhyay 2000; M. Mahdavi 2008; Pacheco 2005).

2.5 Summary

To summarize, there is a need for an assessment model of customer based brand equity that utilizes big data techniques through mining and analyzing brand-related information from online social networks. Few of the prior models introduced in literature harness big data techniques in real time data from social media as their focus was mainly on assessing customer based brand equity with traditional measurement techniques which are costly to obtain and have a certain lag. This study aims to contribute to the literature in the following ways. We contend that traditional methods of data gathering provide a static and sometimes skewed indication of customer perceptions, whereas 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. Second, this research 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. The next sections describe the methodology and experiments made to validate the introduced model.

Chapter 3

Collection and Classification of Data from Online Social Networks

In this section we intend to build on current approaches of mining and analyzing data from online social networks. In particular we introduce a computational framework that assists researchers in (a) mining online social media data-streams in order to detect trends or patterns of consumer behavior and (b) analyze the relationships of data points in order to detect future trends or patterns.

3.1 Introduction

Our intention was to develop a framework that allows researchers to test hypotheses or validate new theories by exploiting the enormous amounts of real-time data-driven information. As we move along the Petabyte era of user generated content, data analysis techniques should accept the challenge of dealing with the implicitly acquired volumes of user data. To this end, our computational framework supports a multitude of data analytic oriented sub-systems that are organized into a layered architecture as we describe in the following section.

The development of a powerful computational tool, however, does not suffice for the efficient manipulation of the enormous data volumes generated within digital social networks. It is essential to formulate an encompassing theoretical framework

for the underlying data analysis tasks, particularly focused on tackling the problems relating to natural language processing, topic modeling, sentiment analysis and semantically – aware community detection. Otherwise stated, our research develops a consolidated framework that detects the underlying causes driving social sentiment. The proposed framework combines topic and sentiment detection approaches to elicit the influential subjects that govern the generation process of consumer attitudes towards brands. This is a fundamental shift from existing analytics perspectives. Instead of estimating social sentiment based on low-level online social verbatim features (e.g. frequency of words with negative or positive sentiment polarity) we propose a high-level abstraction computational model which estimates social sentiment based on the central discussion topics on online social networks. These topics are in turn weighted against their positive or negative influence on the formulation of social sentiment for a given phenomenon, thus providing explanatory insights on what actually drives customer satisfaction.

3.2 Background

Performing Social Media Analytics in the context of a classification exercise needs to address the problems of (a) transforming a large dataset into a set of algorithmically tractable vectors, (b) measuring the sentiment expressed in the data, and (c) identifying the main topics of discussion (Blei, Ng, and Jordan 2003; Zeng et al. 2010; Melville, Sindhwani, and Lawrence 2009) . These problems can be addressed by state-of-the-art machine learning methods, namely corpus vectorization techniques, support vector machines, and probabilistic topic modeling respectively.

3.2.1 Corpus Vectorization

A fundamental prerequisite in order to perform sentiment analysis through the exploitation of any machine learning algorithm is to obtain a mathematical representation of 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), which was originally introduced in Salton et al (1975). The main idea behind VSM is to transform each document d into a vector containing only the words that belong to the document and their frequency by utilizing the so called "bag of words" representation. According to VSM, each document is represented exclusively by the words it contains by tokenizing sentences into elementary term (word) elements losing the associated punctuation, order and grammar information. The underlying mathematical abstraction imposed by VSM entails a mapping which transforms the original purified document to its corresponding bag of terms representation. This transformation can be formulated by the following equation:

$$\varphi: d \to \varphi(d) = [tf(t_1, d), tf(t_2, d), \dots, tf(t_M, d)] \in \mathbb{R}^M \ (\mathbf{1}),$$

where $tf(t_i, d_j)$ is the normalized frequency of term t_i in document d_j given by the following equation:

$$tf(t_i, d_j) = \frac{f(t_i, d_j)}{\max\{f(t, d_j) : t \in d_j\}} \ (\mathbf{2}),$$

given that $f(t_i, d_j)$ is the absolute frequency term t_i in document d_j. Based on the adopted mathematical formulation for the fundamental notions of corpus and dictionary, such that a corpus D of n documents and a dictionary *T* of *M* terms may be represented according to

$$D = \{d_1, d_2, \dots, d_n\} \ (\mathbf{3})$$

and

$$T = \{t_1, t_2, \dots, t_M\} \ (\mathbf{4}).$$

Having in mind Eq.1 and the formal definitions for the notions of corpus and dictionary, the mathematical representation for corpus in the context of VSM can be

done through the utilization of the document-term matrix given by the following equation:

$$D = \begin{bmatrix} tf(t_1, d_1) & \cdots & tf(t_M, d_1) \\ \vdots & \ddots & \vdots \\ tf(t_1, d_n) & \cdots & tf(t_M, d_n) \end{bmatrix} (\mathbf{5}),$$

where N, is typically, quite large resulting in a sparse VSM representation such that a few matrix entries are non-zero. In our approach, in order to mitigate the effect relating to the compete loss of context information around a term, we incorporate the term-frequency inverse document frequency (tf–idf) weighting scheme according to which each term t_i is assigned a weight of the form:

$$w_i = idf(t_i, D) = \log \frac{|D|}{|\{d \in D : t_i \in d\}|} \ (\mathbf{6}),$$

so that the relative importance of each term for the given corpus is taken into consideration.

3.2.2 Support Vector Machines

SVMs are non-linear classifiers that were initially formulated by Vapnik (2013) operating in higher-dimensional vector spaces than the original feature space of the given dataset. Letting $S = \{(\vec{x_i}, y_i) \in \mathbb{R}^n \times \{-1, +1\}, \forall i \in [m]\}$ be the set of m training patterns with associated binary labels, such that -1 denotes the class of negative sentiment and +1 the class of positive sentiment, the learning phase of the SVMs involve solving the following quadratic optimization problem:

$$\min_{\vec{w}, \vec{\xi}, b} \frac{1}{2} \|\vec{w}\|^2 + C \sum_{i=1}^{m} \xi_i \ (\mathbf{7.1})$$

$$s.t. \ y_i(\langle \vec{w}, \vec{x_i} \rangle + b) \geq 1 - \xi_i, \forall i \in [m] \ (\mathbf{7.2})$$

$$\text{and } \xi_i \geq 0, \forall i \in [m] (\mathbf{7.3}).$$

Eqs. 7.1, 7.2 and 7.3 define the Primal optimization problem whose corresponding dual gives rise to a discrimination function of the form:

$$g(\vec{x}) = \sum_{i \in SV}^{m} \alpha_i^* y_i \langle \vec{x}, \vec{x_i} \rangle + b^* \ (\mathbf{8}),$$

where $\{\alpha_i^*, i \in [m]\}$ and b^* denote the optimal solutions for the corresponding optimization variables and SV is the subset of training patterns associated with positive Lagrange multipliers. Given that the training patterns appear only in dot product terms of the form $\langle \vec{x_i}, \vec{x} \rangle$, a positive definite kernel function such as $K(\vec{u}, \vec{v}) = \Phi(\vec{u})\Phi(\vec{v})$ can be employed in order to implicitly map the input feature space into a higher-dimensional vector space and compute the dot product.

3.2.3 Probabilistic Topic Modeling

Probabilistic topic modeling approaches (Blei 2012) 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. 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, the corpus may be once again viewed as a collection D of n documents according to Eq.3 such that, where each document $d \in D$ is a collection words. Therefore, the generative model, provided by LDA, aims at describing the underlying procedure according to which, each document obtains its words. Initially, let's assume the knowledge of T topic distributions for our dataset corresponding to T multinomials containing V elements each, where V stands for the number of terms in our corpus. Therefore, the corpus is composed by a set of unique words that form a vocabulary indexed by $\{1, \dots, V\}$. It is very important to note that the set of terms identified by the LDA topic modeling algorithm are in general different than those identified by the VSM. In the context of LDA, the formal definition of a topic coincides with a probability distribution over a fixed

vocabulary of terms. These topics, in particular, are assumed to be specified before any data has been generated. Subsequently, for each document in the collection, the corresponding words are generated through the utilization of the two-stage process that is described by Blei (2012) below:

1. Randomly choose a distribution over topics.
2. For each word in the document:
    a. Randomly choose a topic from the distribution over topics in step 1.
    b. Randomly choose a word from the corresponding distribution over the vocabulary.

The goal of this exercise is to automatically discover the topics from a collection of words in the corpus.

3.3 Mining Consumer Insights with SMA

Applying hard computational tasks, as described in the previous section, to reveal consumer insights towards brands, depends on a number of steps and prerequisites which should be introduced prior to discussing them in detail in the next chapter. Our proposed model aims at assessing consumer insights towards brands by addressing three fundamental customer based brand equity dimensions. As illustrated in figure 3-1, external brand communications are measured through volume of tweets per topic, per time period. Brand awareness is measured through overall sentiment per day, while brand meaning is depicted through overall sentiment per topic / per day, drilling down to detailed measures of specific events.

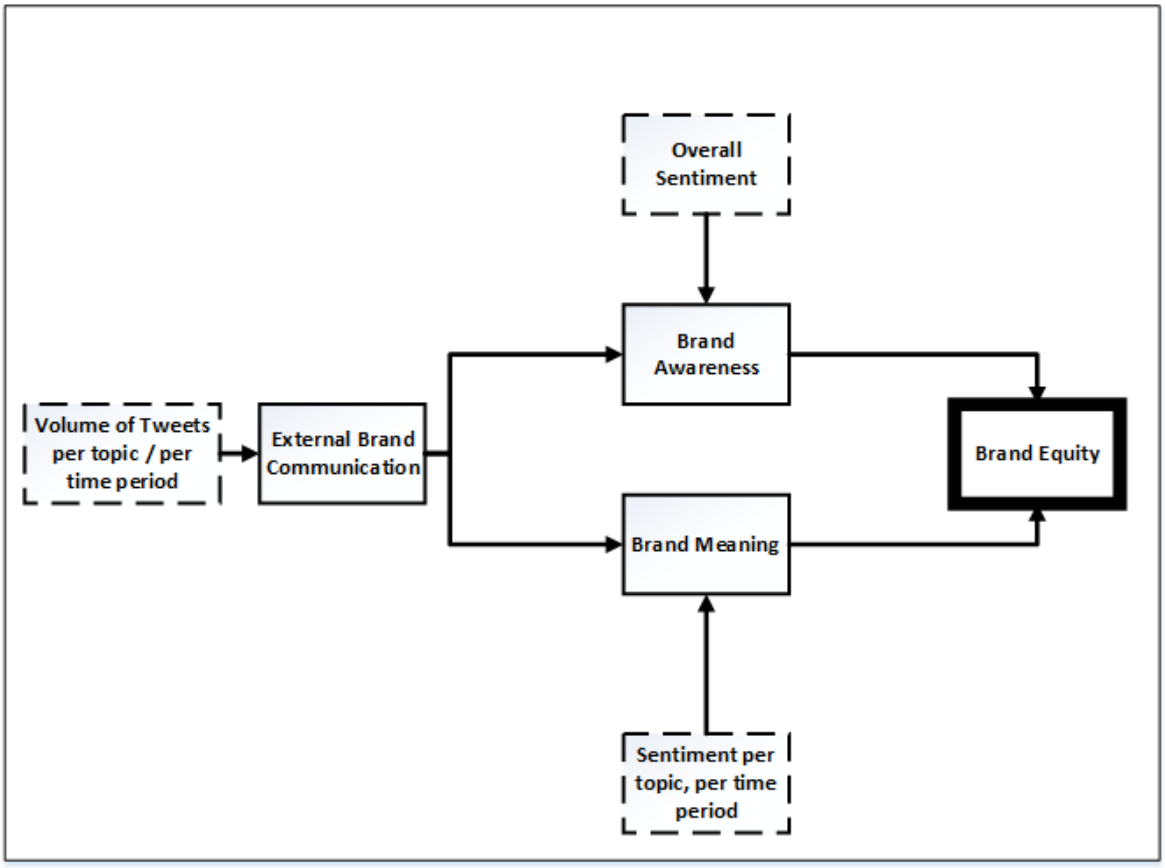


Figure 3.1- Brand Equity Assessment

Our proposed model is based on two disciplines, Marketing and Information Technology. These two fundamental fields should jointly collaborate in order to achieve the organizational strategy objective – in our case customer based brand equity. Each building block includes a number of elements which are illustrated in Figure 3-2. The framework is valid when all elements are combined and can be used to measure the specific organizational strategic objective initially set. In the next section we describe each fundamental element in detail and introduce our cross-section organizational framework for measuring customer based brand equity via SMA.

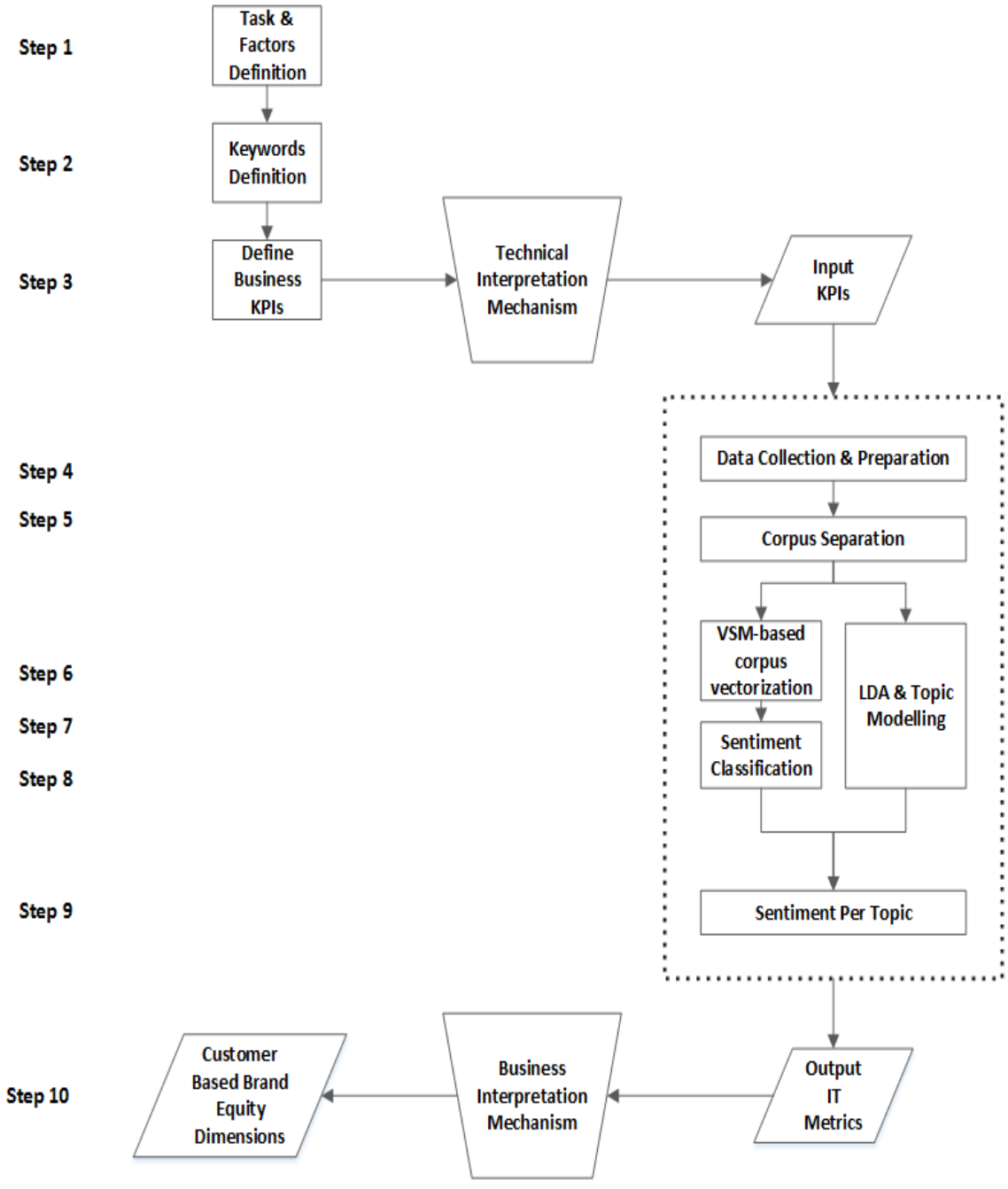


Figure 3.2- Proposed Method

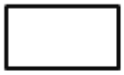
Flowchart Shape	Meaning
	Indicates a process (human or computational driven)
	Indicates an interpretation process led by a person in the organization
	Indicates Data or Metrics

Figure 3.3- Flowchart shape index

3.4 Proposed Method

3.4.1 Step 1: Define Task and Factors to monitor

Organizational actions are and should be measured based on specific factors and actions when evaluating the success and ROI of a given marketing campaign. Being able to quantify customer sentiment towards a brand for a specific campaign requires careful definition of the issue to be monitored, in a given environment.

Customers interact with brands in numerous ways during time and are exposed to different brand messages through various mediums, channels and interactions. In order to explicitly infer customer satisfaction through topic clustering and sentiment tendencies for a specific organizational action, the organization should define in detail the factors it wishes to monitor. Examples could be the launch of a new TV advertisement, a launch of a new online campaign through Twitter, a sponsored event or even a set of outbound campaign actions through different means, during a specific time period. Defining explicitly the task to be monitored and the factors affecting the specific marketing campaign reduces the chance of generating biased results throughout the use of the framework.

3.4.2 Step 2: Define Keywords

Defining specific keywords that describe the brand and characterize the campaign executed are of most 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 case of services) as well as keywords that describe the campaign executed or population targeted. These keywords are used during the data collection and preparation steps in order to form a correct corpus of the objective data set. As described in the next sections, data collected, pass through a series of data cleaning techniques, such as stemming, stop word removal and tokenization. Prior to running this module, if keywords haven't been explicitly defined, the organization risks missing out on valuable information.

3.4.3 Step 3: Define Business KPIs, Interpretation and Input

Every marketing objective consists of specific attributes that define the outcome. Regardless the objective to be measured, the framework will not work unless the outcome measure is broken down to specific measurable components.

The most important step in the framework is to interpret the qualitative measures that define the business objective, to quantitative measures that can be derived from applying machine learning techniques. Cross collaboration between departments as well as presence of Data Scientists within the organizations, contribute towards achieving this task. Our framework aims at revealing insights for three marketing components. External brand communications are inferred through volume of tweets per topic, per time period. Brand Awareness is indicated through Overall Sentiment per Day, while brand meaning is depicted through Overall Sentiment per topic / per day, drilling down to detailed measures of specific events.

3.4.4 Step 4: Data Collection & Preparation

Defining the social media channel to monitor and extract data is the primary decision that the organization should take. Data are collected through the utilization of the relevant API for the given social media channel. The second step involves the precise definition of the time period during data collection. This is a

strategic decision and should be defined depending on the desired outcome. Finally, once data have been collected should subsequently be submitted to a series of data clearing and pre-processing operations. The data preparation process, in particular, involves text tokenization into words, elimination of English stop-words and words with less than three characters, and stem extraction from each word. Therefore, the final version of the corpus will be formed by a collection of purified documents where each document contains the text from a single tweet.

3.4.5 Step 5: Corpus Separation

This step involves the corpus separation procedure that is necessary so that each set of documents concentrates on the particular brands that are being defined. In this context, corpuses for more than one brand should be separated into distinct set of documents. For example if the corpus consists of two brands then it should be separated into two distinct set of documents such that $D = D_A \cup D_V$, where D_A is the set of documents focusing on brand A and D_V is the set of documents concentrate on brand B. Letting n_A and n_B be the number of documents for brand A and brand B respectively, then each corpus can be formally defined according to the following equations:

$$D_A = \{d_1^A, d_2^A, \dots, d_{n_A}^A\} (\mathbf{9})$$

$$D_V = \{d_1^V, d_2^V, \dots, d_{n_V}^V\} (\mathbf{10})$$

3.4.6 Step 6: VSM-based Corpus Vectorization

This step of our method focuses on the corpus vectorization process that is explained in detail in the technical background section. 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. In other words, the exploitation of VSM transforms each set of documents into a corresponding set of feature vectors by utilizing the mapping defined in Eq.1 according to the following equations:

$$\Phi_A = \varphi(D_A) = \{\varphi_1^A, \varphi_2^A, \dots, \varphi_{n_A}^A\} (\mathbf{11})$$

$$\Phi_V = \varphi(D_V) = \{\varphi_1^V, \varphi_2^V, \dots, \varphi_{n_V}^V\} (\mathbf{12})$$

where $\varphi_j^A \in \mathbb{R}^M, \forall j \in [n_A]$, and $\varphi_j^V \in \mathbb{R}^M, \forall j \in [n_V]$.

3.4.7 Step 7: Sentiment Classification

This step encompasses the sentiment classification process which can be further divided into the corresponding training and testing stages. 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, negative or neutral, based on a business 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 feature vector with a raw sentiment value that can be subsequently cast to one of the previously defined classes of positive, neutral and negative tweets. The raw sentiment values associated with each document can be obtained through the utilization of Eq.8 according to the following equations:

$$s_j^A = g(\varphi_j^A), \forall j \in [n_A] (\mathbf{13})$$
$$s_j^V = g(\varphi_j^V), \forall j \in [n_V] (\mathbf{14})$$

such that $-|s_0| \leq s_j^c \leq +|s_0|, \forall c \in \{A, V\}, \forall j \in [n_c]$.

The transformation of the raw sentiment values appearing in Eqs.13 and 14 into the corresponding class identifiers $\{-1,0,+1\}$ is given by the following equation:

$$f(s) = \begin{cases} -1, s < +|s_{thresh}| \\ 0, |s| \leq |s_{thresh}| \\ +1, s > |s_{thresh}| \end{cases} \quad (\mathbf{15})$$

3.4.8 Step 8: LDA – Topic Modeling

This step commits to the topic modeling procedure that is the subject of the technical background section. The LDA probabilistic topic modeling algorithm, in particular, besides unraveling the latent topic structure of each corpus, lays the foundations for an alternative vectorized corpus representation. That is, according to LDA, the documents in each corpus can be treated as points into two T-dimensional probability vector spaces P_A and P_V for the brands under scope. Formally, the application of LDA in each corpus defines a mapping of the following form:

$$\psi_c: D_c \rightarrow P_c, \forall c \in \{A, V\} \ (\mathbf{16})$$

where each document $d_j^c \in D_c, \forall c \in \{A, V\}, \forall j \in [n_c]$ is mapped to a point $\psi_j^c = \psi_c(d_j^c) \in \mathbb{R}^T, \forall c \in \{A, V\}, \forall j \in [n_c]$, such that:

$$\sum_{t=1}^{T} \psi_j^c = 1, \forall c \in \{A, V\}, \forall j \in [n_c] \ (\mathbf{17})$$

The previous definitions imply that the set of documents in each corpus acquire an alternative vector representation according to the following equations:

$$\Psi_A = \psi_A(D_A) = \{\psi_1^A, \psi_2^A, \dots, \psi_{n_A}^A\} (\mathbf{18})$$

$$\Psi_V = \psi_V(D_V) = \{\psi_1^V, \psi_2^V, \dots, \psi_{n_V}^V\} (\mathbf{19})$$

3.4.9 Step 9: Sentiment per Topic

This step lies within the core of our approach since this is where the actual unification of topic modeling and sentiment analysis is performed. The fundamental aim pursued within the execution of this step is to cluster the LDA-based vectors

$\{\psi_j^c\}$ of each corpus into T disjoint sets $\left\{\Psi_c^{\{t\}}\right\}$ according to Eq. 20, so that the entropic measure define in Eq. 21 is minimized. The minimization of Eq. 21 leads to the formation of semantically coherent clusters such that the points pertaining to each cluster exhibit minimum entropic deviation from the corresponding cluster center $\hat{\psi}_t^c$.The quantification of the entropic deviation around the cluster centre is given by the Jensen Divergence measure defined in Eq. 22. The novelty of our approach lies upon the selection of the clustering objectives defined by Eqs. 21 and 22, giving rise to the formation of topically focused groups of tweets which are distributed around cluster centroids that accumulate the vast majority of corresponding topic probability mass on a single topic. In this way, it is possible to semantically associate each cluster of tweets with a unique topic and therefore obtain the sentiment distribution on that particular topic by acquiring the raw sentiment values of the posts pertaining to that specific cluster.

$$C_c = \bigcup_{t=1}^{T}\left\{\Psi_c^{(t)}\right\}, \forall c \in \{A, V\}, \forall t \in [T] \text{ such that } C_c = \Psi_c, \Psi_c^{(k)} \cap \Psi_c^{(l)} = \emptyset, \forall c \in \{A, V\}, \forall k \neq l \quad (\mathbf{20})$$

$$J_{entropic}\left(C_c, \zeta(\cdot,\cdot)\right) = \sum_{t=1}^{T}\sum_{\psi \in \Psi_c^{(t)}} \zeta\left(\psi, \hat{\psi}_t^c\right) \quad (\mathbf{21})$$

$$\zeta(\boldsymbol{u}, \boldsymbol{v}) = \sum_{t}^{T} u_t log_2 \frac{2u_t}{u_t + v_t} + v_t log_2 \frac{2v_t}{u_t + v_t} \quad (\mathbf{22})$$

3.4.10 Step 10: Output IT Metrics & Business Interpretation

The final step is a cross section collaborative action between departments and could be described as the reverse process of step 3. IT metrics produced should be interpreted in the way initially defined and should produce actual quantitative measures of the key components. These quantitative measures should be then assessed by the marketing department in order to produce the final metric for

customer based brand equity, according to their initial criteria. Insights from such an exercise are of outmost importance when evaluating marketing campaigns. The outcome of this exercise could result in a loop by refining initial input, starting again from step 1.

3.5 Summary

In this chapter we introduced a theoretical model which describes the appropriate process of gathering and analyzing data from online social networks. The framework relies on a set of state-of-the-art machine learning techniques which describe the main steps involved in the process of data procession, topic modeling and sentiment classification. We posit that this framework extends extant studies in the field of sentiment analytics, which primarily classify online social networks text entries into positive, negative, and neutral sentiment categories. Instead, we couple sentiment analytics with a topic detection method in order to probe for discussion themes which are highly influential to the formulation of positive and negative opinions. This framework will serve us the necessary technical background to conduct a series of experiments and test the validity of our theoretical predictions regarding assessment of CBBE in the next steps of this research.

Chapter 4

Case Study I: Mining Twitter for Customer Satisfaction in the North American Telecommunications industry

This chapter documents the process and findings of a real world case study, with primary intent to test the efficiency of the framework introduced in chapter 3. In the next sections we briefly discuss the context, scope, objective and rationale for choosing the specific dataset and proceed with presenting the study in detail.

4.1 Study Background

The case study discussed in this Chapter was part of a wider research project called "Sociomine", funded 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). The "Sociomine" project in particular aimed at "*investigating the conditions under which electronic social media favor asymptotic learning or herding and also mine real-time streams of data that are constantly created in electronic social media to infer group behavior and associate it to exogenous economic and social phenomena*". In line with Sociomine's second objective this case study aimed at addressing the challenges of data mining

in real-time social media information streams and correlating public mood with social / economic phenomena.

In particular the case study aimed at presenting results for two of the leading Telecommunication providers in the US, through a four tier layer architecture as illustrated in Figure 4.1, which follows steps 4-9 introduced in the previous chapter. The first layer corresponded to the Real Time Social Network Data Collection Module, performing keyword filtering on the official streaming API of Twitter. This layer was responsible for the real-time data collection process, focusing on gathering tweets that were explicitly referring to a particular subject.

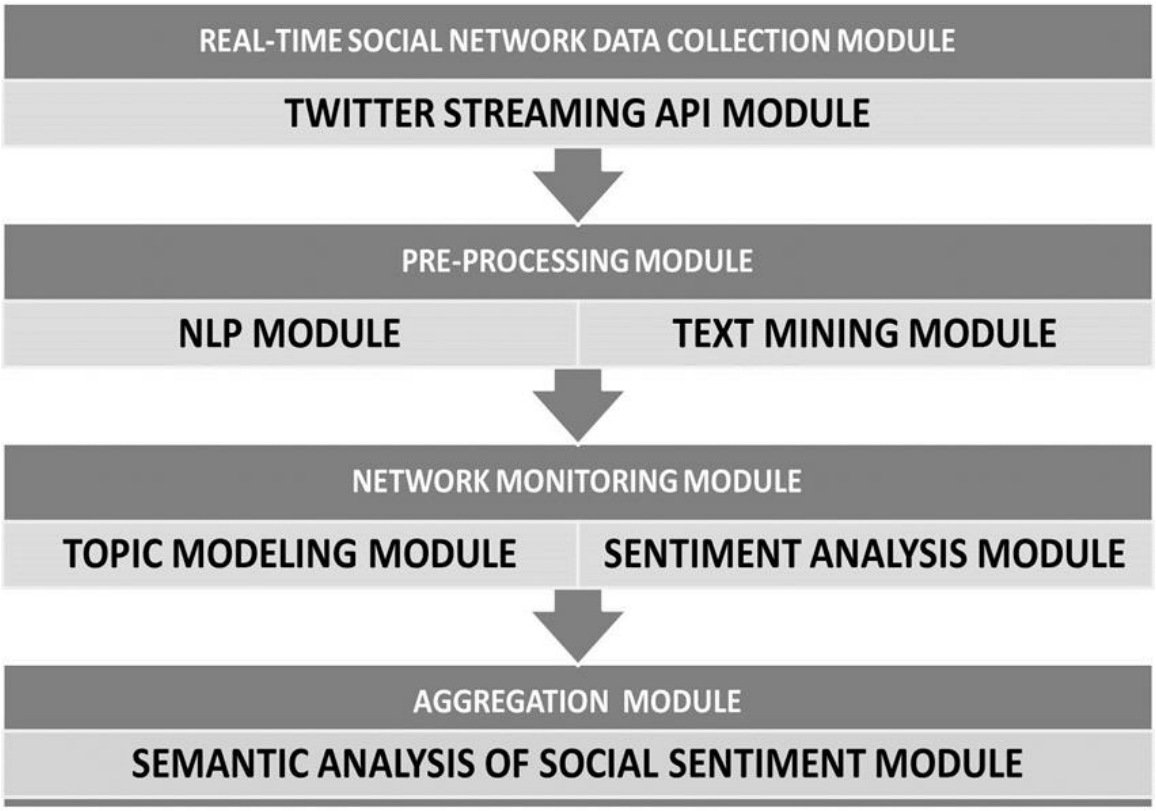


Figure 4.1- Framework Architecture

The second layer, referred to the Pre-Processing Module, assigned with the task of data preparation by exploiting the sub-modules of Natural Language Processing (NLP) and Text Mining.

The data preparation process involved text tokenization into words, elimination of stop-words and words with less than three characters, and stem extraction from each word. The previous sequence of operations can be summarized as the corpus formation process.

The Network Monitoring Module performed the actual tasks of Topic Modeling and Sentiment Analysis. Probabilistic topic modeling was performed through the utilization of the Latent Dirichlet Allocation (LDA) technique (Blei, Ng, and Jordan 2003). Sentiment Analysis, on the other hand, was conducted by taking advantage of a state-of-the-art classifier, namely Support Vector Machines (SVMs). SVMs are non-linear classifiers that were initially formulated by Vapnik in (1995), operating in higher-dimensional vector spaces than the original feature space of the given dataset.

Finally, the fourth layer of the proposed framework (Aggregation Module) assigned the task of combining the results of the previous processing levels. Our primary objective was to exploit a high-level abstraction computational model which estimates social sentiment based on the central discussion topics on online social networks.

4.1.1 The North American Telecommunications Sector

The North American Telecommunications sector is one of the leading mobile broadband sectors worldwide, representing increasingly important revenue opportunities for mobile operators. Taking into consideration that the market is being saturated and revenue from new subscriptions is increasingly deteriorating, mobile carriers tend to focus on customer service and high levels of customer satisfaction, in order to retain registered customers and maintain a low churn rate.

4.1.2 AT&T vs. Verizon: Financial Figures in a Saturated Environment

According to AT&T's 2012 fourth quarter results, published on January 24, 2013, AT&T posted a net increase in total wireless subscribers of 1.1 million in the fourth quarter to reach 107.0 million subscribers in service with annual operating revenue of $127 billion. Verizon is number two in retail connections, with 98.2 million subscribers in service and $75.9 billion annual revenue in 2012.

According to a study [8] conducted on February 13, 2012 Total U.S. Telecommunications Industry Revenues reached $985 billion during 2010 with Annualized Total Wireless Service Revenues matching $159.9 billion. Arguably, the telecommunications sector, and the wireless services group in particular, is one of the leading drivers in the U.S economy, rendering competition between carriers very intense.

Based on a qualitative survey, of director-level and above marketing and business executives responsible for retention strategies at 40 service providers across North America, Europe, Asia Pacific and Central and Latin America, conducted from June to July 2011 by Amdocs[9], 66% of operators believed that customers are less loyal today than they were two years ago, 70% of service providers cited customer retention and loyalty as the critical factor for driving growth, with a strategic marketing prioritization shift from customer acquisition and market share to long-term customer engagement. Due to market saturation and increasing competition, 82% of service providers said that customer loyalty programs would be "very important" or "important" over the next five years to their company's strategy.

---

[8] http://www.columbia.edu/cu/consultingclub/Resources/Telecommunications_Pablo_PrietoMunoz.pdf [Accessed: 29-July-2016]

[9] http://www.amdocs.com/news/pages/amdocs-customer-retention-and-loyalty.aspx [Accessed: 29-July-2016]

4.2 Scope & Objectives

In this case study, we try to examine if mobile wireless carriers can benefit from applying the model introduced in the previous chapter, in order to discover insights from consumer perceptions in social media.

In order to examine this question we focus on two aspects:

- First we apply our model to data gathered from Twitter, of customers mentioning the two leading mobile wireless carriers in North America (AT&T and Verizon) and examine if CBBE can be assessed through analyzing and monitoring social media networks (Twitter in particular).

- Second we compare the results, to results gathered from two customer satisfaction surveys performed through call interviews, by two different independent research laboratories, in order to examine if there is correlation between the results.

4.2.1 Survey Based Customer Satisfaction Results

On January 2013, Vocalabs [10] published the National Customer Service Survey (NCSS) on Mobile Phones based on data collected from 2009 through 2012, through independent research, tracking results for AT&T, Sprint, Verizon and T-Mobile. The study draws on some pretty insightful results which are presented in brief below and along with results from the Forrester study, acted as a benchmark of comparison to the sentiment analysis performed on the collection of tweets.

[10] http://www.vocalabs.com/sites/default/files/NCSS-Mobile-Phone-Q4-2011.pdf [Accessed: 29-July-2016]

The National Customer Service Survey for Mobile Phone Customer Service is a continuous survey beginning July 2009, and data from all four years are presented. Customers were interviewed immediately after a customer service call to one of the companies in the report. The survey measured customer perceptions of the quality of the customer support they received from AT&T, Sprint, T-Mobile, and Verizon.

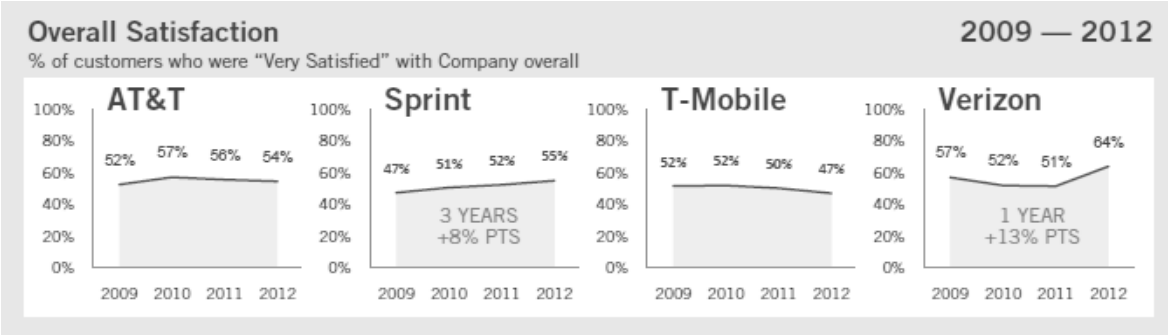


Figure 4.2 Overall customer satisfaction scores from NCSS as provided by Vocalabs on January 2013

The most dramatic trend over the past year is the significant improvement in Verizon's customer satisfaction and loyalty. The company posted a very significant 13-point gain in overall customer satisfaction and a six-point increase in the percentage of customers who would buy another Verizon phone if given the chance.

Apart from the overall customer satisfaction survey in regards to company services, we looked at results examining customer satisfaction for AT&T and Verizon's Wireless services (4G & LTE) as presented by Forrester on the fourth quarter of 2012, as these services are being used by the carriers to attract new customers, focusing on high broadband speed rates as a form of competitive advantage.

According to Forrester[11], both AT&T and Verizon score 66 (ranked as OK on the rating scale).

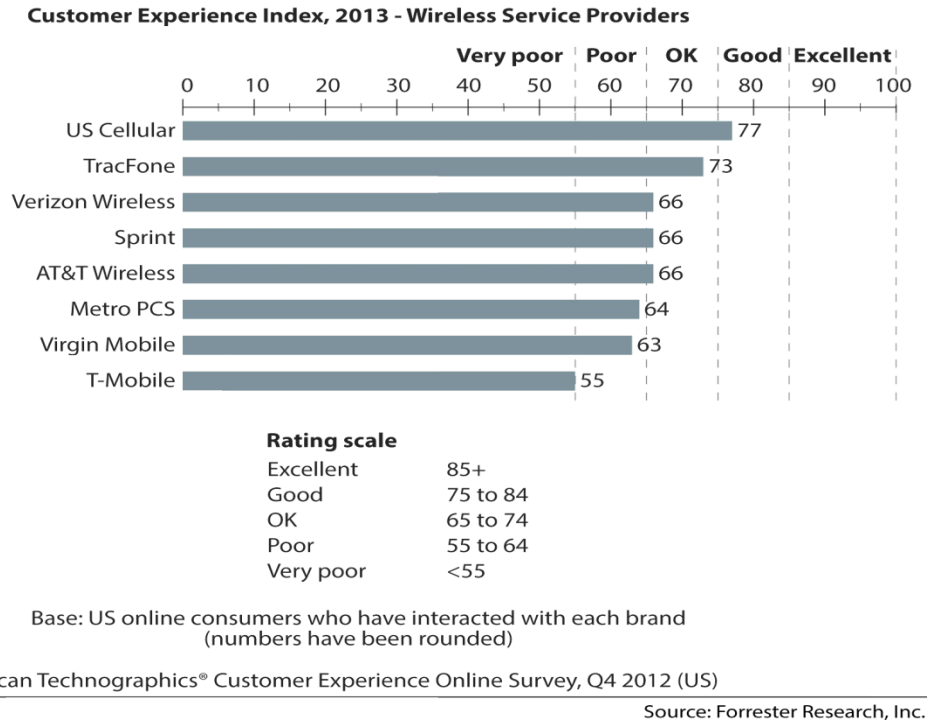


Figure 4.3 Customer satisfaction scores from Customer Experience Index 2013 on Wireless Service Providers.

4.3 Case Study Steps

Steps 1-3: The mobile telecommunications business sector was selected for the application of our framework, focusing on the two leading carriers in the USA, AT&T and Verizon. Taking into consideration that the market is being saturated and revenue from new subscriptions is increasingly deteriorating, mobile carriers tend to focus on customer service and high levels of customer satisfaction in order to maintain a low churn rate. In this context, it is a matter of critical importance to

[11] http://solutions.forrester.com/Global/FileLib/Reports/The_Customer_Experience_Index_2013.pdf [Accessed: 29-July-2016]

be able to measure the overall customer satisfaction level of their respective brand, through measures such as customer based brand equity.

The main reason behind the choice of the specific industry is the rapid rise of mobile use through smart devices, which gives the subscribers the ability to have instant access to SMN, any time anywhere, 24/7. Smart device users now have the power to express opinions through SMN at the exact moment of customer-brand interaction providing data that, if analyzed appropriately, could provide valuable insights about customer based brand perception. Whereas customer responses through questionnaires or face to face surveys provide a rather more formal view of customer's perceptions, customers responses via SMN, tend to reflect real emotion and feeling towards a brand, at the exact point of interaction, be that negative or positive.

It is within this context that providing a framework for measuring customer satisfaction and being able to classify customer feedback as positive or negative has a significant place in organizational actions. Whereas previous studies have addressed measurement through qualitative techniques, such as scale-based measurements, we argue that customer satisfaction can be measured through a quantitative approach by classifying customer sentiment through the use of SMA.

Steps 4-5: We collected and analyzed a set of over 135,000 tweets during the time period between February 2nd and February 26th, 2013, by utilizing the Streaming API of Twitter. The specific period was chosen as carriers were on their peak of advertising campaigns as well as heavy voice and data usage, due to Super Bowl XLVII, the most watched annual sporting event in the world[12]. Prices in marketing campaigns have increased every year, with advertisers paying as much as $3.5

[12] Nielsen Media Research, Statistics on Super Bowl TV Viewership in the US, http://www.statista.com/statistics/216526/super-bowl-us-tv-viewership/ February 2013. Accessed on 9/12/2013.

million for a thirty-second spot during Super Bowl XLVI in 2012 [13] , while a segment of the audience tunes into the Super Bowl solely to view commercials[14]. The data collection process was focused on gathering tweets that were explicitly referring to the two leading mobile broadband carriers AT&T and Verizon. This task was accomplished by parsing the official streaming API of Twitter through keyword filtering on the terms "AT&T" and "Verizon". Twitter was chosen as the SMN medium, for gathering data as studies show that it's the fastest growing social platform in the world[15] and more Twitter users chose to tweet from mobile devices rather than PCs[16]. The resulting dataset contained a total number of 66,000 and 70,000 tweets for AT&T and Verizon respectively, which was subsequently submitted to a series of data clearing and pre-processing operations. The data preparation process, in particular, involved text tokenization into words, elimination of English stop-words and words with less than three characters, and stem extraction from each word. Therefore, the final version of our corpus was formed by a collection of purified documents where each document contained the text from a single tweet.

Steps 6-7: The LDA topic modeling technique was applied on the two of n_A=65,971 and n_V=70,576 purified tweets, by setting to T = 10 the number of topics to be extracted. Table 4.1 summarizes the ten topics that were extracted for the corpus associated with AT&T, while Table 4.2 summarizes the ten topics that were

---

[13] "30-Second Television Ads During Super Bowl XLVI to Cost $3.5 Million | TIME.com." 2016. http://business.time.com/2012/01/23/the-super-bowl-has-morphed-into-an-entire-season-for-advertising/, Accessed on 9/12/2013.

[14] Kotala, C., Commercials as big as game, Florida Today, http://pqasb.pqarchiver.com/floridatoday/doc/239286633.html, Accessed on 9/12/2013.

[15] Global Web Index Blog, Twitter Now The Fastest Growing Social Platform In The World, http://blog.globalwebindex.net/twitter-now-the-fastest-growing-social-platform-in-the-world/ , Accessed on 9/12/2013.

[16] Strategy Analytics, Twitter Users Are Switching From PCs to Tablets and Phones,http://www.strategyanalytics.com/default.aspx?mod=pressreleaseviewer&a0=5350, Accessed on 9/12/2013

extracted for the corpus associated with Verizon. Each line consists of the nine most indexed words in each cluster of tweets, while the 10th column in each row (omitted for appearance purposes) relates to the brand.

1	phone	got	get	iphone	need	dont	service	shit	anyone
2	galaxy	note	iphone	samsung	battery	cordless	for	update	blackberry
3	commercial	the	new	that	little	girl	talking	kid	boy
4	jobs	job	sales	retail	consultant	time	support	part	manager
5	iphone	verizon	sprint	mobile	one	htc	free	case	apple
6	network	lte	2012	from	video	years	invested	and	speeds
7	lte	like	apple	car	iphone	black	new	que	16gb
8	commercials	kids	little	the	love	funny	these	lol	those
9	park	service	f***	verse	wireless	verizon	center	san	today
10	sucks	internet	time	get	service	day	work	lol	like

Table 4.1- AT&T LDA-BASED TOPICS

1	wireless	case	iphone	att	phone	sales	cover	jobs	skin
2	iphone	sprint	att	hate	service	like	mobile	sucks	lol
3	att	contract	time	stop	high	commercial	tell	internet	wireless
4	fios	data	internet	unlimited	service	day	wifi	comcast	verizons
5	apple	iphone	nokia	ipad	black	lumia	16gb	best	att
6	motorola	droid	verizons	apps	android	razr	battery	wireless	vodafone
7	htc	droid	dna	phone	update	one	wireless	android	galaxy
8	phone	get	got	new	dont	need	lol	anyone	store
9	center	f***	blackberry	show	washington	why	tickets	the	wizards
10	wireless	galaxy	samsung	lte	center	the	att	android	new

Table 4.2-Verizon LDA-BASED TOPICS

Figures 4.4 & 4.5 present the results obtained by applying our clustering approach on the corpuses of tweets concerning AT&T and Verizon, given that the number of clusters to be formed was equal to the number of topics, T = 10, generated by the LDA topic modeling technique. The clustering results indicate that the adapted entropic measure-based clustering objective has the ability to form semantically coherent clusters that are distributed around topically focused cluster centroids. The figure illustrates the corresponding cluster centroids for AT&T and Verizon, where it is evident that each cluster centroid accumulates the majority of the associated topic probability mass on a single topic. It is this clustering result that

allows us to infer the semantic structure of the corpuses under investigation since each cluster of tweets concentrates around a single-topic centroid. It is of primary importance to note that our clustering approach has the ability to generate T = 10 clusters of tweets that are distributed around the exactly T = 10 semantic axes spanning the complete set of tweets for both telecommunication companies. The results of the LDA topic modeling indicate three prevailing factors as a result of Twitter buzz during the Super Bowl period for both brands, such as Product Portfolio, Sponsored Events and Advertisement actions.

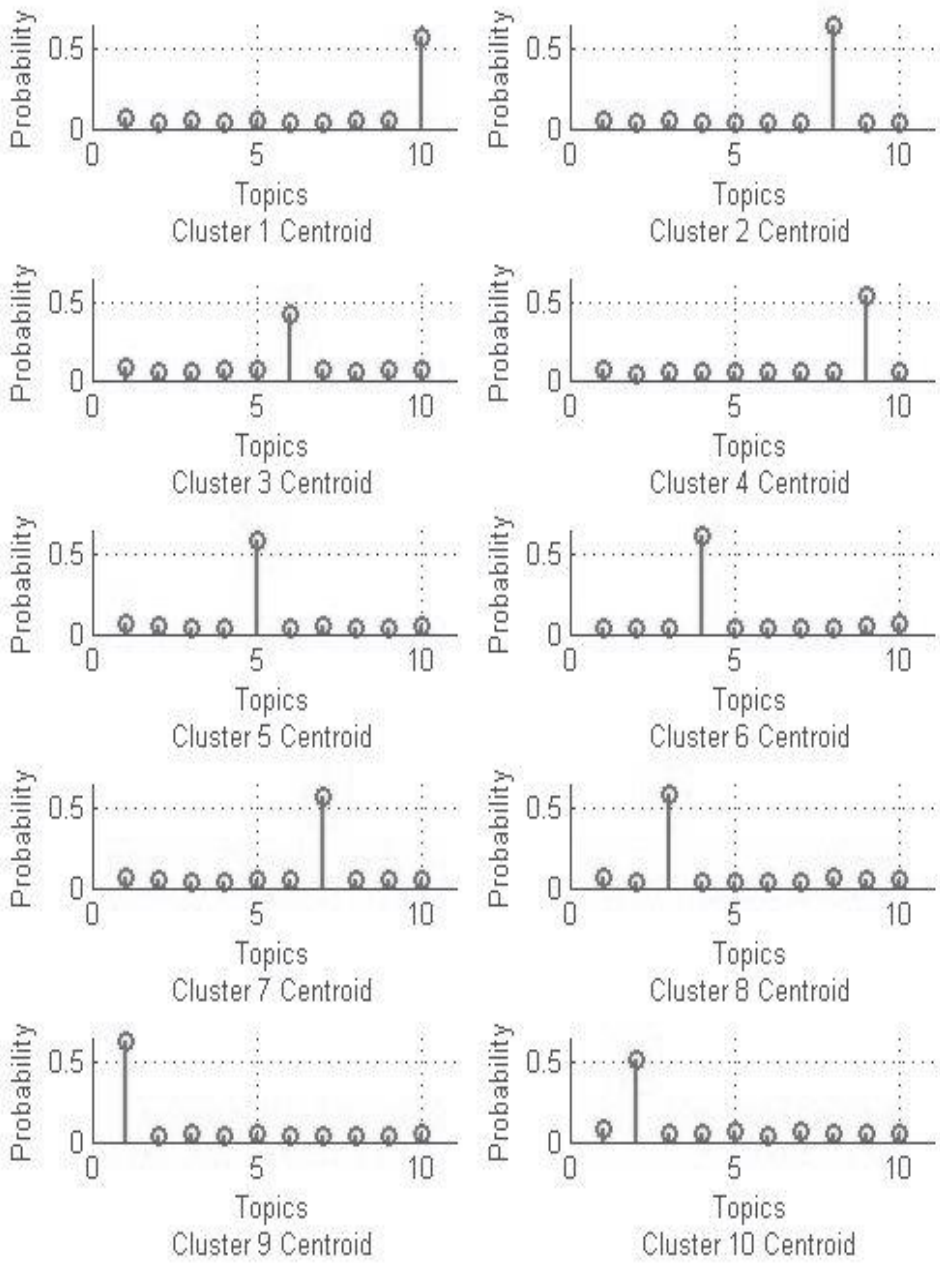


Figure 4.4: AT&T Cluster Centroids

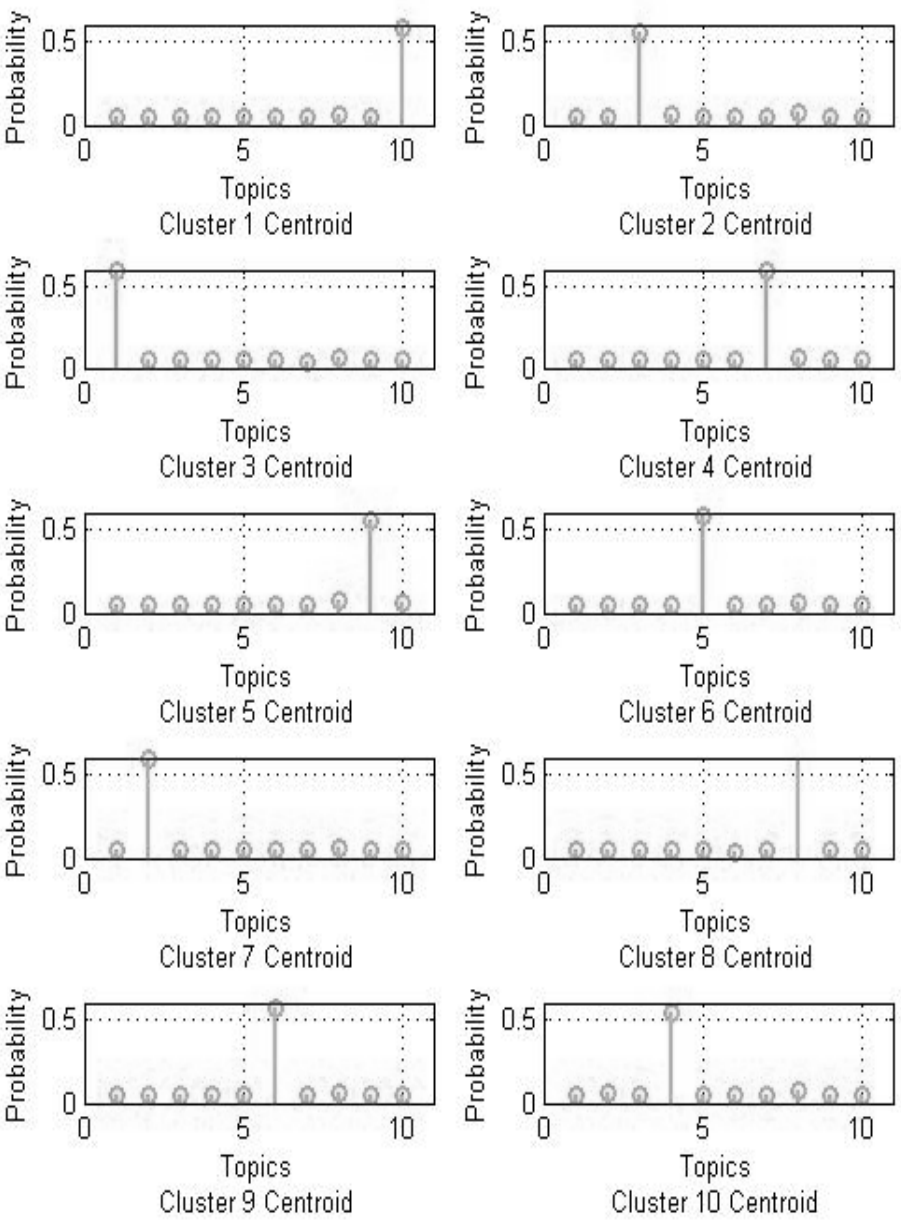


Figure 4.5: Verizon Cluster Centroids

Steps 8-9: Applying our framework in such a large amount of tweets required a reasonable amount of labeled data (i.e. tweets already classified as positive, negative or neutral, based on a business perspective classification). This ensured that the SVM algorithm ran with accuracy, providing robust results that limit the amount of fault. These labeled data where in turn used by the SVM algorithm as a benchmark, in order to score the number of tweets that are in scope of the sentiment exercise. A group of eight undergraduate students were employed to manually label a set of 7,223 collected tweets, in terms of sentiment, as positive (1), neutral (0) or negative (-1). A sample of tweets for AT&T and Verizon can be found in Appendix 1. Nationality and base of residence of the students (Greece) ensured bias towards a given company, as all eight students had never used any of the companies services and seven out of eight students didn't have previous experience of brand exposure to AT&T and Verizon. Example tweets of each category are presented below in tables4. 4 to 4.5, while the results of this exercise are depicted in tables 4.6 and 4.7.

RT @xxxxxx: AT&T is the best.	1
Just got great service from AT&T over the phone thats never happened with any other phone company #HappyHappyHappy	1
Does anyone else think those AT&T commercials with the little kids are funny and cute or is it just me?	1

Table 4.3 - Sample of Tweets marked as positive

At the AT&T Store with my Homegirl! Next stop, StoneCrest Mall!	0
@zzzzzs: I just saw @xxxxx name on the iPhone 4S screensaver! AT&T store?	0
Do yall pay attention to these AT&T commercials?	0

Table 4.4 - Sample of Tweets marked as neutral

I dont hate kids but those AT&T commercials makes me hate them	-1
Apple and AT&T is pissing me off.	-1
AT&T why u hate me	-1

Table 4.5 - Sample of Tweets marked as negative

Total tweets assessed: 2,939	
Negative (-1)	901
Neutral (0)	1,113
Positive (1)	925

Table 4.6 – AT&T Labeled Tweets

Total tweets assessed: 4,284	
Negative (-1)	1,684
Neutral (0)	1,450
Positive (1)	1,150

Table 4.7 – Verizon Labeled Tweets

In order to test the accuracy and validity of the SVM algorithm on the sentiment classification problem, the total amount of labeled tweets (7,223 Tweets) was split into a 95% training data - 5% testing data ratio. The percentage of tweets, which are labeled data that have already been classified as positive or negative, formed the set of training data. On the other hand, the percentage of tweets to be scored by the SVM algorithm formed the set of testing data. This procedure produced a series of confusion matrices letting us compare how accurately the SVM algorithm classified the testing data in accordance to our already classified labeled data. The results for AT&T and Verizon on this subset of data are presented in tables 4.8 to 4.11.

AT&T Results (Training)		
Accuracy = 0.914697		
Actual / **Prediction**	**Positive**	**Negative**
Positive	800	56
Negative	92	787

Table 4.8- AT&T Sentiment Classification Confusion Matrix on Training Data

AT&T Results (Testing)		
Accuracy = 0.923077		
Actual / **Prediction**	**Positive**	**Negative**
Positive	44	1
Negative	6	40

Table 4.9-AT&T Sentiment Classification Confusion Matrix on Testing Data

Verizon Results (Training)		
Accuracy = 0.897141		
Actual / **Prediction**	**Positive**	**Negative**
Positive	1484	116
Negative	161	932

Table 4.8-Verizon Sentiment Classification Confusion Matrix on Training Data

Verizon Results (Testing)		
Accuracy = 0.638298		
Actual / **Prediction**	**Positive**	**Negative**
Positive	63	21
Negative	30	27

Table 4.10-Verizon Sentiment Classification Confusion Matrix on Testing Data

The results of the exercise proved that the SVM algorithm could achieve a very good testing accuracy percentage allowing us to run the SVM algorithm for the full data set with a threshold value set to 0.3, classifying all collected tweets according to their sentiment (see table 10). With the results in hand we performed in depth statistical analysis to produce the three IT metrics that when combined, reveal customer satisfaction and conducted a sample technical to business interpretation allowing us to present them as actionable business results.

Full Data Set of Tweets Data Results	
AT& T Results	**Verizon Results**
Minimum decision value: -2.078615	**Minimum decision value:** -2.320534
Maximum decision value: 2.424328	**Maximum decision value:** 2.090736
Absolute Threshold Value 0.3	**Absolute Threshold Value** 0.3
Decision Value Based Estimated	**Decision Value Based Estimated**

Negative Patterns 9244	**Negative Patterns** 9946
Decision Value Based Estimated Positive Patterns 6819	**Decision Value Based Estimated Positive Patterns** 6641
Decision Value Based Estimated Neutral Patterns 49908	**Decision Value Based Estimated Neutral Patterns** 53989

Table 4.11- Sentiment Classification Results

Step 10: The first metric concerns volume of tweets, distributed across topic and time period specified. Figures 4.6 and 4.7 illustrate the evolution of twitter buzz for AT&T and Verizon and indicate a rather high but stable number of tweets in regards to topic 10 for AT&T and topic 8 for Verizon which indicates that a significant amount of tweets will be generated for both brands regardless of specific marketing actions due to the relatively high brand equity that both companies possess in the market. What is interesting is to focus on topics which show a sudden change in volume during time period span. AT&T has a significant spike in volume for topics 3 and 8 from time period 8 to 11. A closer look at the topic reveals that the factor closely associated to the topic is the buzz around a TV commercial launched during the specific time period.

The second metric concerns the calculation of the sentiment classification of each brand per time period. The corresponding results for AT&T and Verizon are presented in figure 4.8, according to which AT&T starts receiving positive sentiment towards its brand during the period identified in the previous section as the launch of a TV advertisement. This metric combined with the total amount of tweets generated for each brand, as presented previously reveal positive brand awareness for AT&T in contrast to negative brand awareness for Verizon during the same time period.

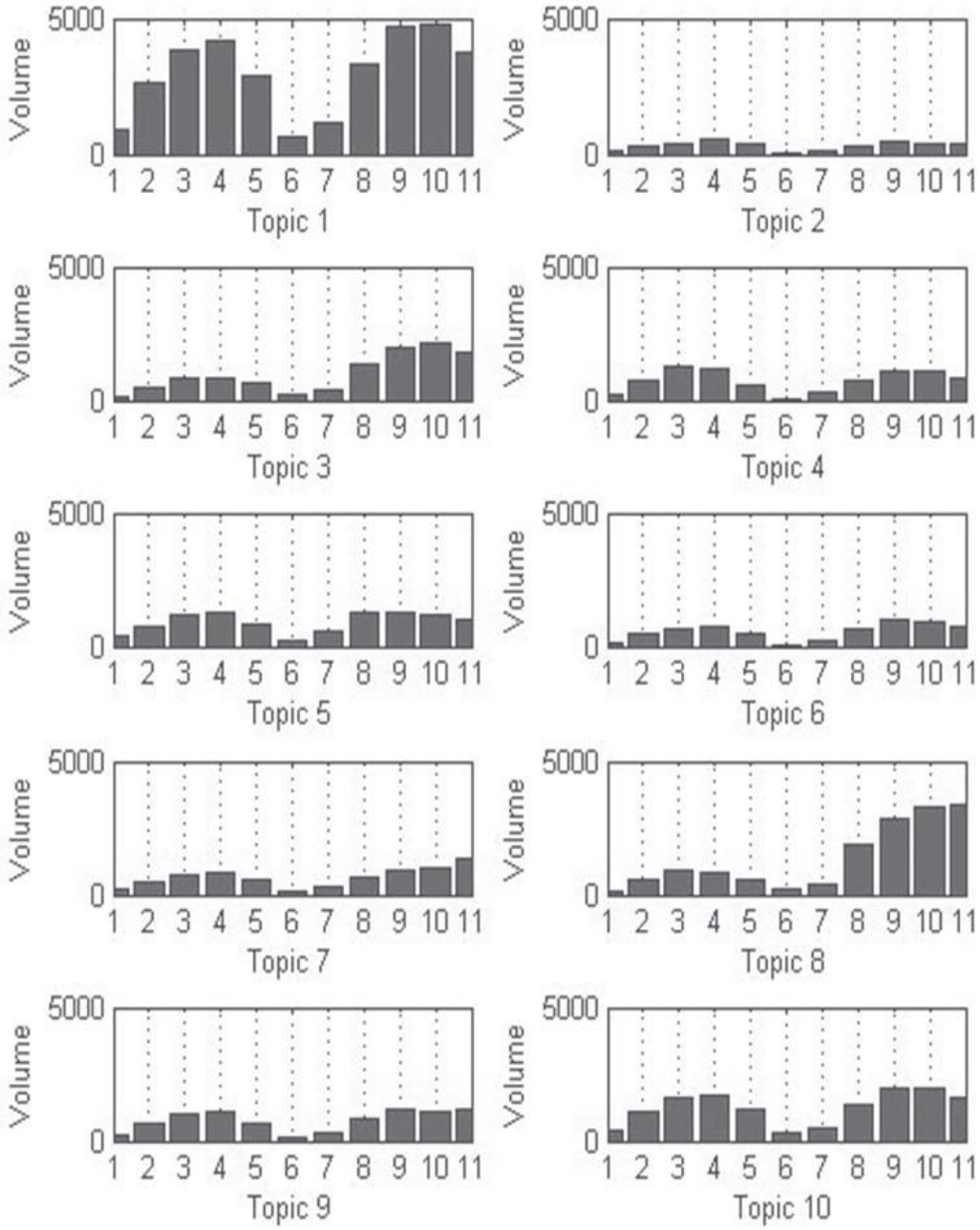


Figure 4.6: AT&T Volume per topic / per time period

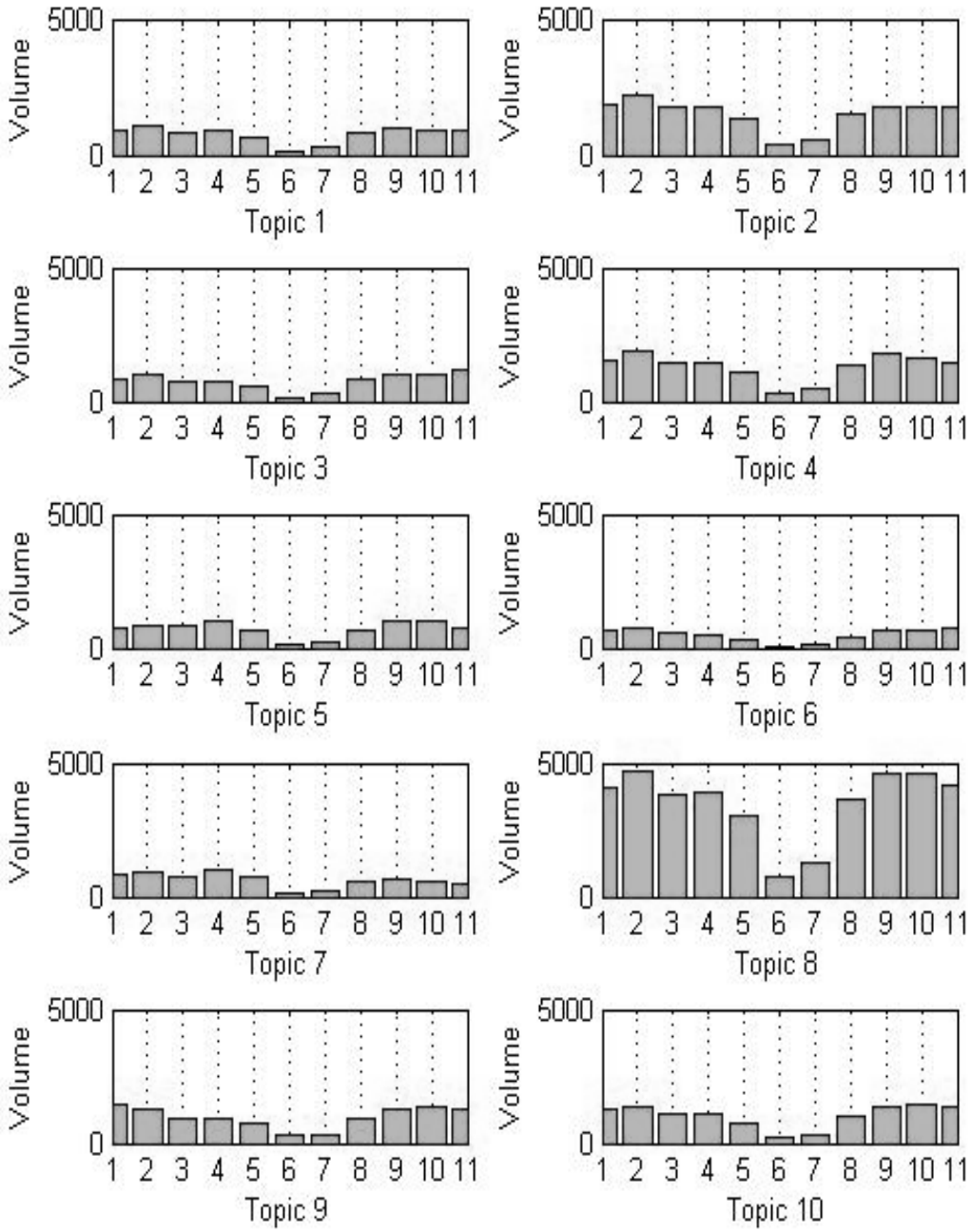


Figure 4.7: Verizon Volume per topic / per time period

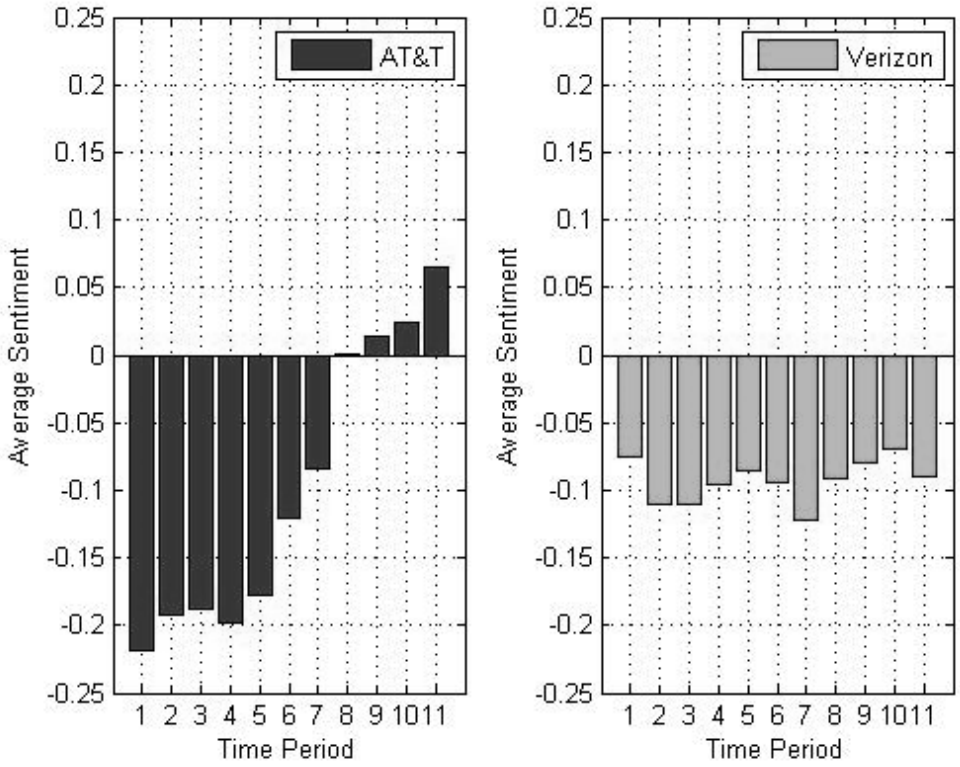


Figure 4.8: Overall Sentiment

Figures 4.9 and 4.10 illustrate the evolution of the average public sentiment per time period for each topic. It is evident that the fraction of positive customer attitude towards AT&T can be uniquely attributed to the average positive sentiment of topics 3 and 8, while the fraction of positive customer attitude towards Verizon can be uniquely attributed to the average positive sentiment of topics 1, 5 and 10. Under the light of these associations it is easy to deduce that the fraction of positive brand meaning towards AT&T can be exclusively attributed to the advertisements released by the company during the period under investigation. It is important to note that for the case of Verizon brand meaning associated with the product portfolio is not exclusively positive since there are clusters associated with the same semantic factor that accumulate a negative sentiment value on the average.

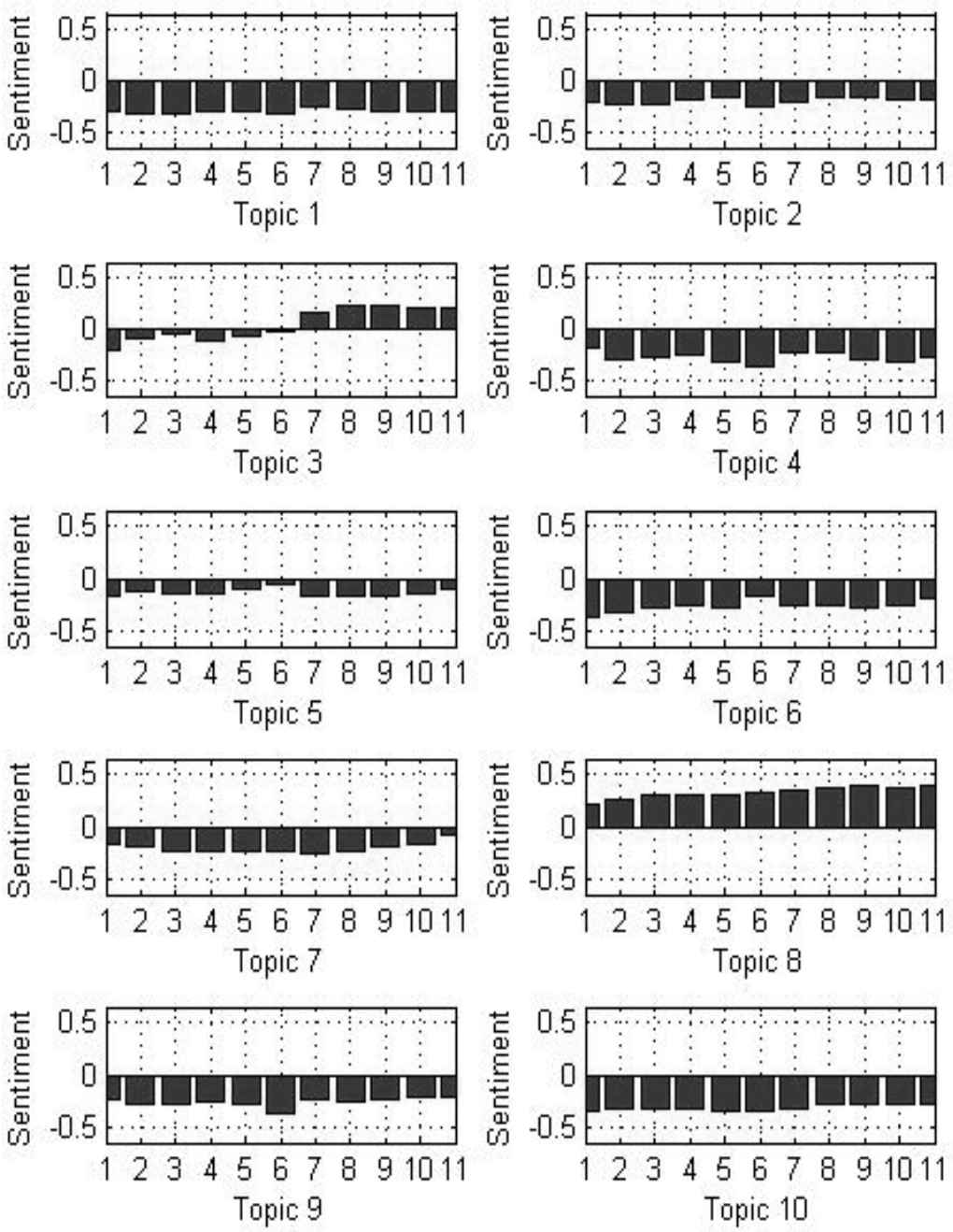


Figure 4.9: AT&T Sentiment per topic / per time period

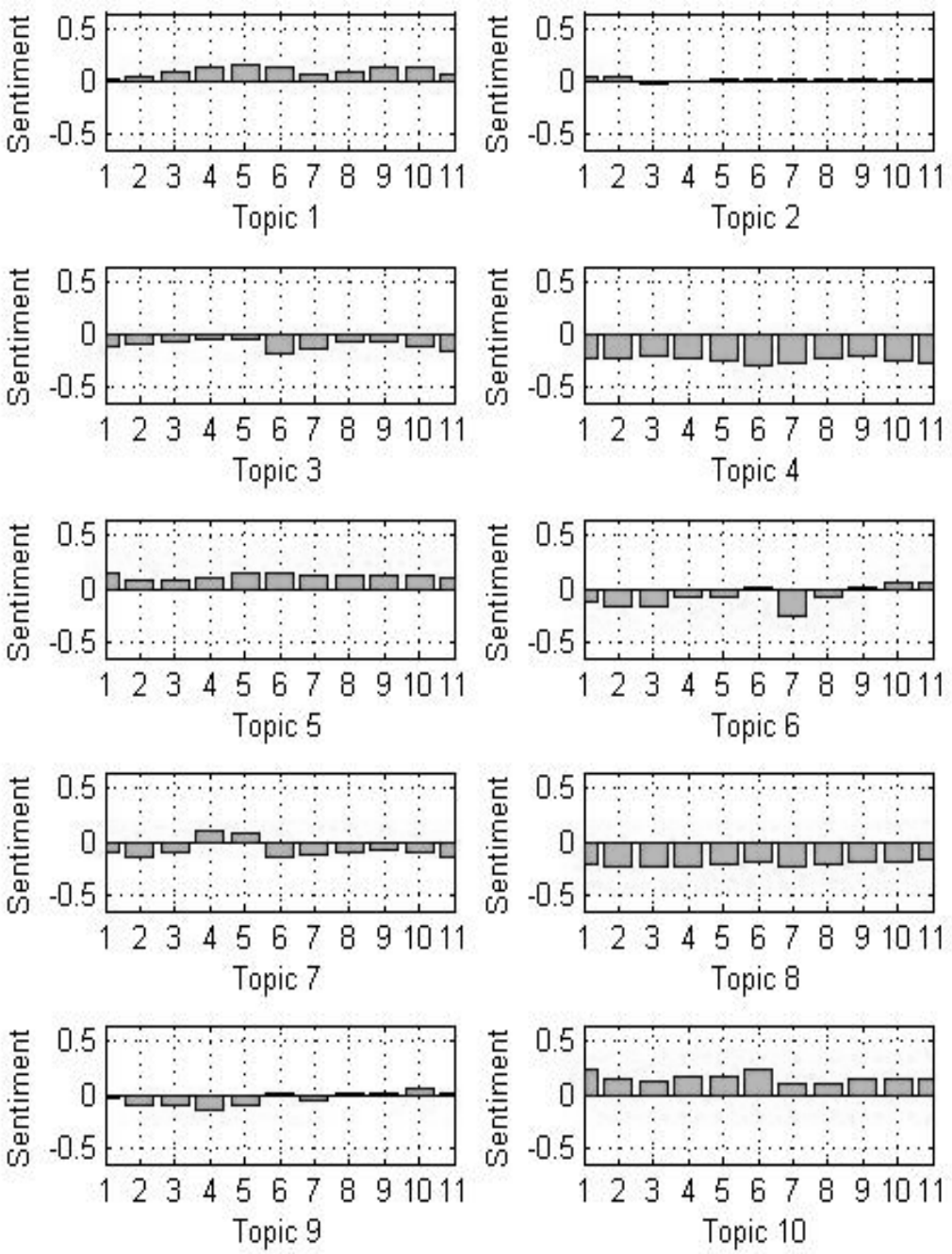


Figure 4.10: Verizon Sentiment per topic / per time period

4.3.1 Comparison with Survey Results

From the results derived above, we can calculate customer satisfaction as measured from Twitter as:

[Decision Value Based Estimated Positive Patterns / (Decision Value Based Estimated Positive Patterns + Decision Value Based Estimated Negative Patterns)] * 100

Thus the figures for AT&T and Verizon are 51% and 30% respectively. The comparison with the surveys presented in the previous section, can be visualized as Figure 4.11 and Figure 4.12 below.

Figure 4.11 Customer satisfaction scores for Verizon.

The results indicate two very important findings. Running SVM algorithm for sentiment detection, in very large sets of data, can prove highly accurate if provided with a good labeled set of already scored data. The algorithm proved the ability to learn and score accurately based on the labeled data provided, given the assumption that the data set is properly cleaned.

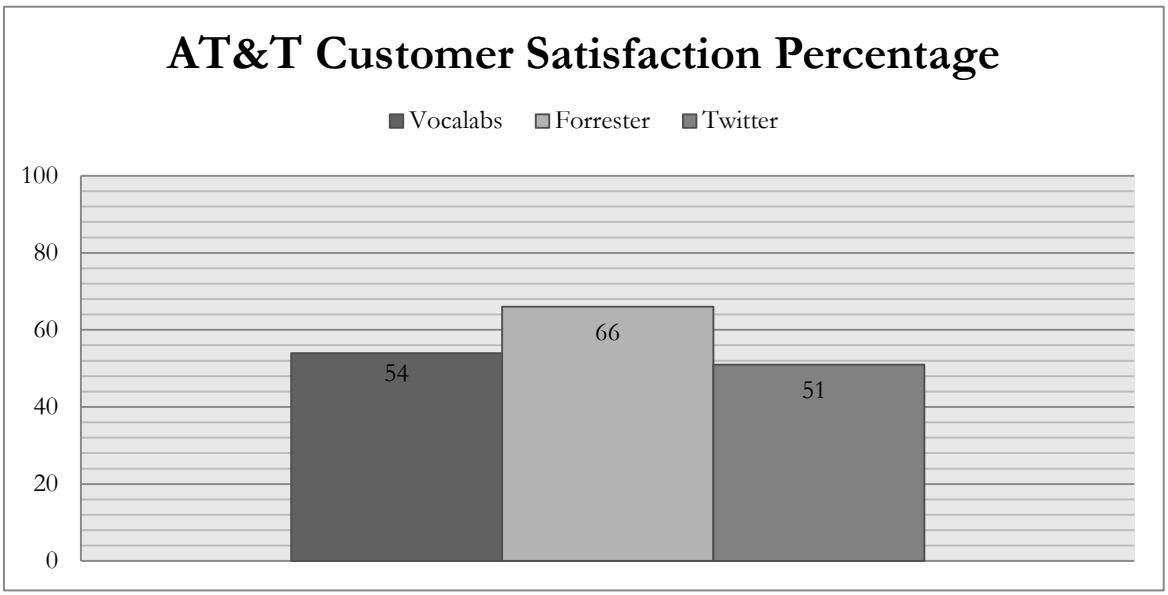


Figure 4.12 Customer satisfaction scores for AT&T.

Second, the results showed that mining twitter for customer satisfaction can prove a very big asset for any organization if used appropriately. Given the nature of the medium, sentiment analysis in sets of data of a specific time frame, can provide useful insights about the specific period. We propose that such exercises be ran during monitored periods that the organization expects high load of conversation to arise in twitter triggered by specific events. In our case, we found that twitter users showed a positive tendency towards a specific commercial AT&T had recently launched, while showing a negative tendency towards Verizon, due to broadband problems that the service was facing for a few hours during the period monitored. This reveals that, although this approach can prove very insightful for drawing conclusions during the specific period, it shouldn't be compared with results from customer satisfaction surveys that the data collection timeframe spans during large periods of time.

4.4 Conclusions

4.4.1 Insights

This case study aimed at examining how the two leading mobile broadband carriers located in the broader North American area, AT&T and Verizon, can benefit from monitoring and performing sentiment analysis, on tweets sent from users in Twitter, mentioning keywords in scope of, or related to the two carriers.

The model was empirically applied to guide the analysis of more than 135,000 tweets in order to estimate customer satisfaction scores for two of the leading mobile broadband carriers in North America: AT&T and Verizon, during Super Bowl XLVI. Our research results showed that data gathered from Twitter, if mined, cleaned and scored appropriately can prove of outmost importance, as this information depicts customer sentiment towards the respective carrier on a real-time and a more intimate or straightforward basis. Mobile broadband carriers will benefit from and improve customer satisfaction if they include such an activity in their customer satisfaction methodology.

We propose that carriers perform such an activity during periods of events that trigger twitter users to actively participate in discussions and express their opinions. These activities could be during a launch of a new commercial, launch of new services, or even disaster situations where wireless services are not responding. The amount of information that could be gathered in such situations in such a small period of time can prove salutary in situations where quick responses may need to be taken in order to maintain the churn rate low.

We conclude that comparing results from customer surveys with results gathered from twitter could prove useful as a benchmark of the validity of the results that are generated through offline telephone conversations but in no means can one method

replace the other. This is due to the difference of nature of each medium used to gather customer opinion and from the authors' perspective both methods should be used complementary.

4.4.2 Limitations

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 all SMN that provide the relevant API for data extraction. Future researcher can consider applying the framework in multiple case studies, in regards to both number and purpose.

A second limitation which is appropriately raised by Z. Tufekci (2013) , is the statement that "Twitter has emerged as a model organism of big data". Results derived from mining this SMN channel for sentiment don't necessarily reflect real world viewpoints, mainly due to the fact of the characteristics of the population choosing to express views from the specific mean. Nevertheless metrics derived from SMN can provide an additional pool of knowledge for organizations that shouldn't be neglected or ignored in any case.

Third, in order to assure generalizability of the results, the framework should be applied in organizations of different verticals and market sizes. We urge future researches wishing to apply the framework, to take this parameter into account in the experimental applications they chose to proceed.

Finally, a dependency rather than limitation that should be stated is the need for appropriate interpretation mechanisms to exist within the organizations wishing to apply the framework. Interpreting and transforming business metrics to IT metrics and vice versa, is still a very intriguing task that needs to be applied scientifically, recently drawing towards data science, in order to ensure robust results that adhere to initial requirements.

4.4.3 A Critique of the Model and Results

Although the results presented in the case study are of particular interest, they raise a series of questions in regards to the model's accuracy and validity.

Results showed that the overall thrust and application of the model was in its way poorly defined, without clearly stating its main goal and contributions. It could be argued that the model and case study presented, tried to accomplish too much (conceptual framework, machine learning / technical contribution, and empirical analysis) and consequently neither of them was sufficiently developed to make a substantive contribution to the field. Furthermore, it was evident that the model needed to be more concise on the domains it was drawing and needed better grounding in the sub disciplines.

Accordingly, the model and case study stated that it could provide timely feedback on a specific campaign or campaigns. In this case, there should have been more specified standards on which tweets to include in the analysis and which not. Taking from the examples provided in the relevant tables, there were several tweets about the person's experience in the AT&T/Verizon store commenting on the service, which had nothing to do with a certain campaign. Including these kinds of tweets when assessing the effectiveness of certain campaigns can bring noises to the results.

One of the major drawbacks of the study was the split of data in to discrete time periods as well as the data and event period the data were collected. The collected tweets were from Feb 2 to Feb 26, 2013. Super Bowl 2013 was on Feb 3 2016, which was almost at the beginning of the data period. However, the effects identified spiked primarily between periods 8 and 11, which was a fairly later time window. In addition, the Super Bowl was a single-day event. Claims that the identified trend over several weeks can be associated with the advertising on a single day could not be further justified.

Another drawback of the case study was the unbalanced manual coding of tweets (2,939 vs. 4,284), which was mainly caused by the unbalanced coding speed of the students performing the task. Considering the total population has approximately 66000 AT&T tweets and 70000 Verizon tweets, AT&T tweets are underrepresented in the labeling sample.

Results also showed a significant drawback by using a keyword-based search query on the Twitter streaming API as this failed to ensure gathering of tweets that relate to the voice-of-the-customer to infer customer satisfaction results. Unrelated tweets were also collected that didn't contribute and made the analysis even harder. Even within the results of the study, certain topics were indeed determined by messages irrelevant to customer satisfaction for the brands under investigation.

Regarding topic modeling; one characteristic of the standard LDA approach is that the number of inherent topics within the document collection must be defined in advance. The number was set to 10, without being tested or evaluated to demonstrate that ten topics indeed lead to the best model.

From the above it is evident that the conceptual model, which is the key contribution of this research, needed to be further developed, grounded in literature and subjected to further evaluation with careful use of keyword level mining and analysis. A natural approach for this could be through a Design Science research approach (von Alan et al. 2004) by positioning the problem under investigation and introducing an IT Artifact that tackles the problem.

Conversely, the theoretical framework and technical steps needed to be much better developed and in particular evaluated. A natural approach to tackle this is by comparing the novelty the algorithms bring with other alternative approaches and demonstrate superior quality.

4.5 Summary

This chapter presented a real world case study, empirically applying the theoretical model introduced in chapter 3, to guide the analysis of tweets in order to infer customer insights and estimate customer satisfaction rates for two of the leading mobile broadband carriers in North America: AT&T and Verizon, during Super Bowl XLVI. Drawing on the findings, we were able to investigate how specific marketing insights could be detected from mining data from an online medium such as Twitter, as well as draw on the theoretical framework and test how we can classify such data into discrete topics and semantically coherent groups.

Nonetheless, it was evident that the findings of chapters 2 & 3 sparked additional questions which require a refined model and further experimentation. The findings indicated the need for the model to be further discussed in detail in regards to the marketing construct under investigation and its research approach, the technical contribution and evaluation compared to similar approaches, and further experimentation to test its validity. To this end, the following chapter introduces a refined model entitled *"A Computational Model for Mining Consumer Perceptions in Social Media"*, presented through a Design Science research approach, which describes how fundamental aspects of customer based brand equity can be inferred from mining consumer perceptions from social media. The model is consequently enhanced with a novel Genetic Algorithm, which improves clustering of tweets in semantically coherent groups, and in turn is compared with similar approaches, such as the k-means algorithm.

Chapter 5

A Computational Model for Mining Consumer Perceptions in Social Media

In the previous chapters we identified a need to further enhance the model introduced in chapter 3, present it through a theoretically grounded research approach, focus on presenting results for a specific marketing construct and further evaluate the model by comparing the technical steps included, with existing approaches. As such, this chapter introduces a refined method for eliciting influential subjects that govern brand equity assessment, by mining and analyzing consumer perceptions from online social network data. To do so, it applies a design science research approach. The rest of the chapter describes the design and evaluation of a computational model that lays out a proposed method on how consumer perceptions could be optioned, starting from a marketing perspective, the technical steps necessary to construct the assessment metrics, and concludes with how the results could be interpreted and utilized in brand equity assessment exercises.

5.1 Introduction

As discussed in the previous chapter, the application of our model in the AT&T and Verizon dataset, revealed substantial conceptual, methodological and inferential weaknesses on the research and presentation approach. In its current form, the presentation of the model rather tried to fit too much and created confusion regarding the main research contribution. To overcome such shortcomings, we restructured and enhanced parts of the study, presenting our research through a design science research approach, allowing us to focus on the conceptual contributions of the study and theoretically articulating our model's logic in a much more comprehendible mean.

Before presenting the model it would be expedient to highlight the main changes of the refined approach compared to that introduced in chapter 3. In this chapter we restructure our model, highlighting the computational methodology, by enhancing the technical details, the validation, and the theoretical grounding of the model. The core computational model has been refined and now comprises of 6 steps (Figure 5.1). In particular we proceed with the addition of a novel Genetic Algorithm which improves clustering of tweets in semantically coherent groups, that acts as an essential prerequisite when searching for prevailing topics in big pools of data. This addition allowed us to infer further insights in regards to awareness and meaning consumers place for the given brand. The sentiment classification process has also been modified so that it is performed in a binary classification setting for both the labeled and unlabeled subsets of tweets. In fact, each document is now assigned with a particular sentiment value that corresponds to the soft decision output of the SVM classifier. Therefore, the absolute threshold parameter is no longer relevant. Moreover, decisions concerning the polarity of the sentiment (positive or negative) are internally made by the classifier by taking into consideration the sign of the corresponding soft decision value.

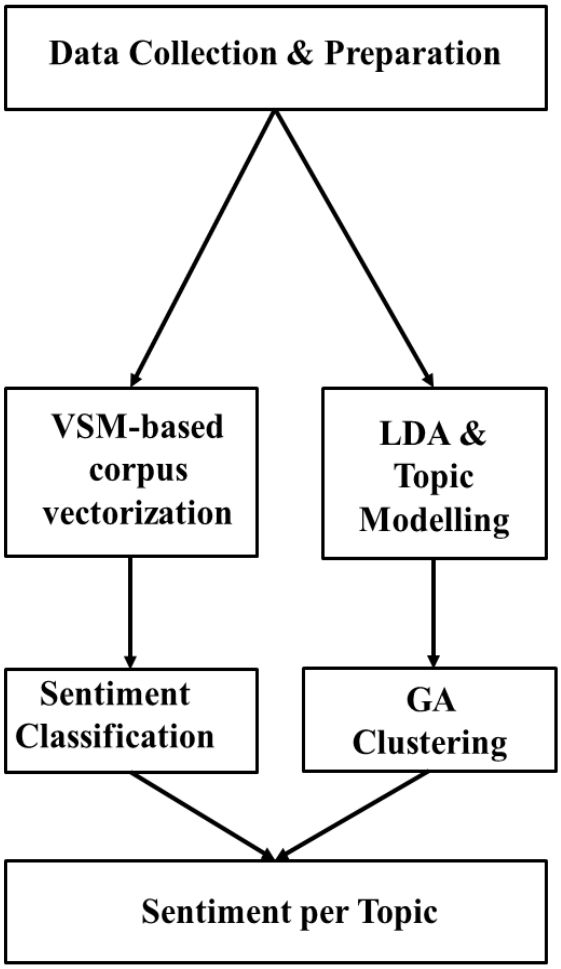


Figure 5.1: Computational Engine

To address the problem of relating indices generated by our computational model to the theoretical domains of CBBE, we have restructured the model which now comprises of 8 steps (rather than 10). Our model generates prevailing topics and clusters in the given corpus, while also producing four key output metrics that reveal consumers perceptions. Metric pairs 1 & 3 and 2 & 4 provide insights on Brand Awareness and Brand Meaning respectively.

We also stress the following points:

- Our model doesn't aim at replacing current methods, but rather stresses the need to complement the overall assessment by not neglecting insights that can be inferred from online social network data.
- We stress the need for organizations to provide keywords related to the brand (i.e. the product name or company name in case of services) as well as keywords that describe the campaign executed or population targeted (in 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 initial objectives set.
- Although positive attributes of brand meaning can be extracted by various indicators, negative attributes can primarily be extracted from customer perceptions through social media data. We thus stress the need to focus on customer generated data from social media which due to the nature of the medium can be mined and analyzed in near real time.
- The model reveals two fundamental dimensions of customer based brand equity from social media, namely brand awareness and brand meaning, inferred from consumer perceptions.

The rest of the chapter introduces the refined method for eliciting influential subjects that govern brand equity assessment, by mining and analyzing consumer perceptions from online social network data, in detail.

5.2 Research Approach

This study follows a design research approach (Gregor and Jones 2007; von Alan et al. 2004) and positions the work according to Gregor & Hevner's framework (2013) following the publication schema as showcased in a paper by McLaren et al (2011). According to March & Smith (1995)

> *"acquiring knowledge concerning both the management of information technology and the use of information technology for managerial and*

organizational purposes involves two complementary but distinct paradigms; behavioral science and design science".

As such, Design Science addresses research through the building and evaluation of artifacts designed to meet the identified business need. The main goal of this research approach is to address important unsolved problems in unique or innovative ways or solve problems in more effective or efficient ways. Our model falls under the improvement quadrant (figure 5.2) in the DSR knowledge contribution framework (Gregor and Hevner 2013) as its main aim is to show how and why this new solution differs from previously presented ones. As pointed in the previous chapters, while customer based 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.

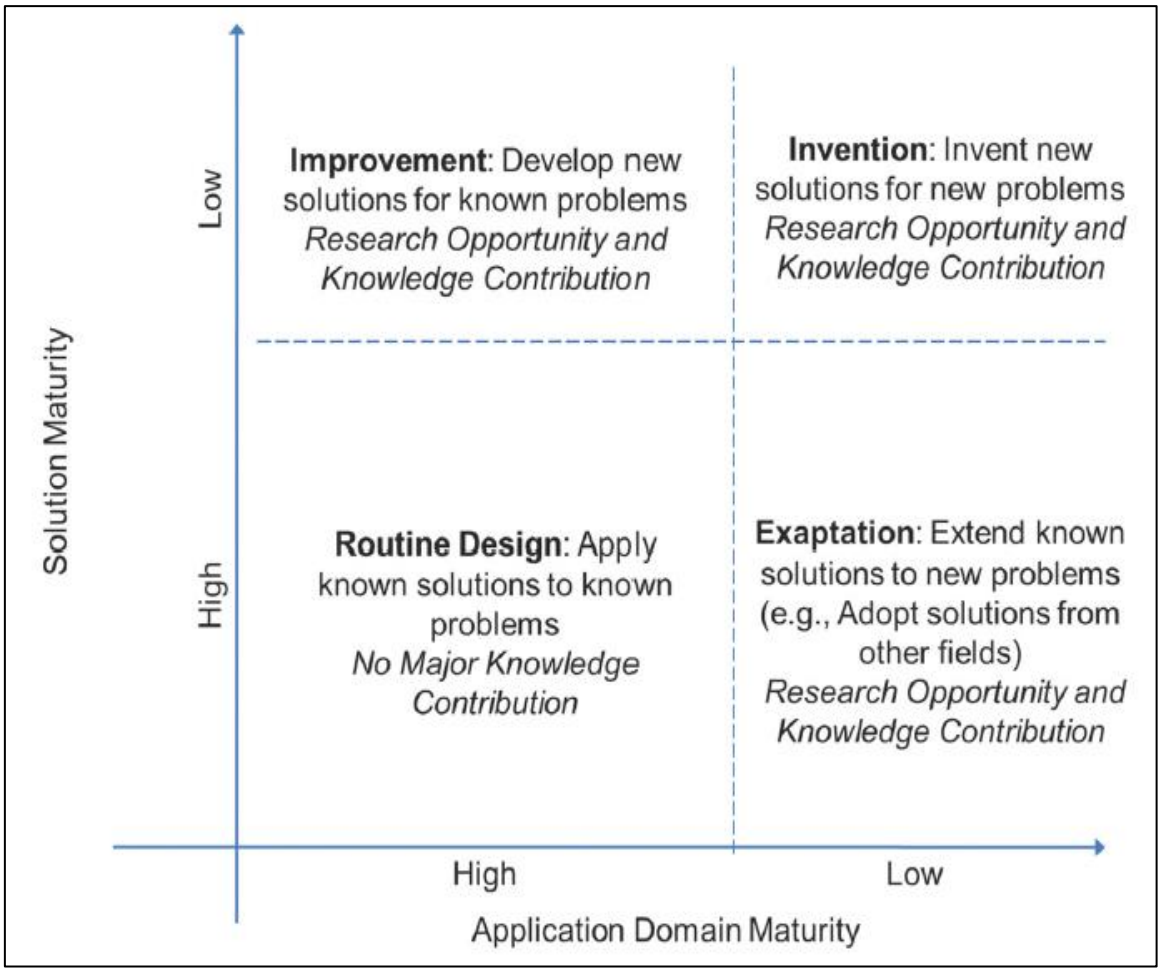


Figure 5.2: DSR knowledge contribution framework (Gregor and Hevner 2013)

Design, implementation and evaluation of our model are justified using prior theory and the case study findings (see next chapter). 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 and business decisions during the assessment lifecycle.

5.3 The model

The novelty of this research lies upon 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.

The proposed model draws on Aaker's definition on Brand Equity (1992), Melville's framework (2009) for assessing marketing constructs through SNA techniques and utilizes Berry's conception (2000) for service brands, using the dimensions of brand awareness and brand meaning. In the next sections we describe the design and evaluation of our computational model that lays out a proposed method on how fundamental dimensions of brand equity can be optioned by mining consumer perceptions from SMN, the computational methods necessary to be followed and conclude with how these results can be interpreted from a marketing perspective through the application of the model in a relevant case study.

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 5.3. In this section we introduce our computational model and we describe each fundamental element in detail.

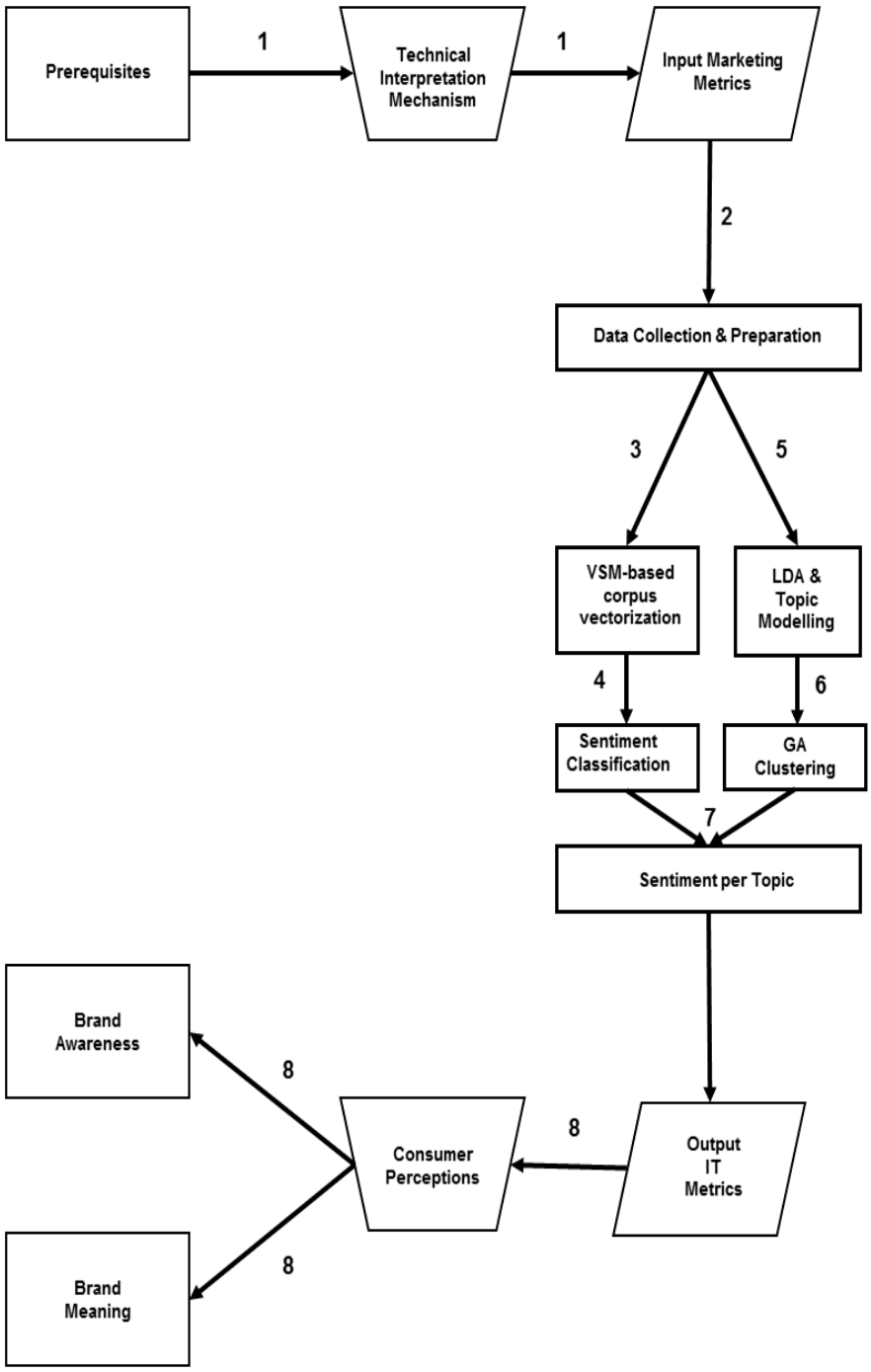


Figure 5.3: The model

5.3.1 Step 1: Prerequisites

Defining specific keywords that describe the brand and characterize the campaign executed are of most 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 case of services) as well as keywords that describe the campaign executed or population targeted (in 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.

5.3.2 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 (Hoffman and Fodor 2010; McDonald, de Chernatony, and Harris 2001), list pros and cons of each medium and the choice should be based on the desired outcome the organization wishes to gain knowledge about. 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 should subsequently be submitted to a series of data clearing and pre-processing operations. The data preparation process, in particular, 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 by a collection of purified documents where each document contains the text from a single tweet, given by the following equation:

$$\boldsymbol{\mathcal{D}} = \{\boldsymbol{d_1}.\boldsymbol{d_2}, \dots, \boldsymbol{d_n}\} \quad (\mathbf{1})$$

where *n* is the number of available documents.

5.3.3 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 at obtaining 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 d 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 terms representation. That is, each document $d \in \boldsymbol{\mathcal{D}}$ will be finally represented through the utilization of an M-terms dictionary

$$\boldsymbol{T} = \{\boldsymbol{t_1}, \boldsymbol{t_2}, \dots, \boldsymbol{t_M}\} \quad (\mathbf{2})$$

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

$$\boldsymbol{\phi}: \boldsymbol{\mathcal{D}} \rightarrow \mathbb{R}^{\boldsymbol{M}} \quad (\mathbf{3})$$

such that:

$$\boldsymbol{\phi(d)} = [\boldsymbol{tf(t_1, d)}, \boldsymbol{tf(t_2, d)}, \dots, \boldsymbol{tf(t_M, d)}] \quad (\mathbf{4})$$

where $tf(t_i,d)$ is the normalized frequency of term t_i in document $d \in \boldsymbol{\mathcal{D}}$ 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 compete loss of context information around a term. In this context, each term t_i is assigned a weight w_i of the following form:

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

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:

$$\boldsymbol{\Phi} = \boldsymbol{\phi}(\boldsymbol{\mathcal{D}}) = [\boldsymbol{\phi_1}, \boldsymbol{\phi_2}, \dots, \boldsymbol{\phi_n}] \quad (\mathbf{6})$$

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

5.3.4 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, in particular, 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:

$$\boldsymbol{\Lambda} = \{\boldsymbol{\lambda_0^*}, \boldsymbol{\lambda_1^*}, \dots, \boldsymbol{\lambda_m^*}\} \quad (\mathbf{7})$$

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 Λ and the associated training feature vectors define a hyper plane within the implicitly induced higher-dimensional feature space which serves as the discrimination boundary between the subspaces of positive and negative tweets defined as:

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

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

$$\boldsymbol{K(\phi, \phi_j) = exp\left(-\gamma \cdot \left|\left|\phi - \phi_j\right|\right|^2\right)} \quad (9)$$

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:

$$\boldsymbol{g}: \mathbb{R}^{\boldsymbol{M}} \rightarrow [\mathbf{0}, \mathbf{1}] \quad (\mathbf{10})$$

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 implicitly 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 ϕ_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:

$$\boldsymbol{S = \{s_1, s_2, \ldots, s_n\}} \quad (\mathbf{11})$$

such that -1 ≤ s_j ≤ +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 which is subsequently utilized in order to estimate the average sentiment value for a given subset of tweets.

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

$$\boldsymbol{\psi}: \boldsymbol{\mathcal{D}} \rightarrow \boldsymbol{\mathcal{P}} \quad (\mathbf{12})$$

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

$$\sum_{t=1}^{T} \boldsymbol{\psi_j}(\boldsymbol{t}) = \mathbf{1}, \quad \forall_{\boldsymbol{j}} \in [\boldsymbol{n}] \quad (\mathbf{13})$$

The previous definitions imply that the set of documents acquires an alternative vector representation according to the following equation:

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

5.3.6 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 $\boldsymbol{\Psi}$, our evolutionary clustering method consists in determining a set $\boldsymbol{\Psi}^*$ of *K* cluster centroids, given as:

$$\boldsymbol{\Psi}^* = \{\boldsymbol{\Psi}_1^*, \ldots, \boldsymbol{\Psi}_K^*\} \quad (15)$$

that implicitly define an optimal *K*-partitioning of $\boldsymbol{\Psi}$ such that:

$$\boldsymbol{\Psi} = \bigcup_{\boldsymbol{i=1}}^{\boldsymbol{K}} \boldsymbol{\Psi_i} \quad (\mathbf{16})$$

and

$$\boldsymbol{\Psi_r} \cap \boldsymbol{\Psi_l} = \emptyset, \forall\, \boldsymbol{r}, \boldsymbol{l} \in [\boldsymbol{K}]: \boldsymbol{r} \neq \boldsymbol{l} \quad (\mathbf{17})$$

where

$$\boldsymbol{\Psi_i} = \{\boldsymbol{\psi} \in \boldsymbol{\Psi}: \left|\left|\boldsymbol{\psi} - \boldsymbol{\Psi_i^*}\right|\right| < \left|\left|\boldsymbol{\psi} - \boldsymbol{\Psi_r^*}\right|\right|, \forall \boldsymbol{r} \in [\boldsymbol{K}]: \boldsymbol{r} \neq \boldsymbol{i}\} \quad (\mathbf{18})$$

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:

$$\min_{\{\boldsymbol{\Psi_1^*},\ldots,\boldsymbol{\Psi_K^*}\} \in = \mathcal{P}^{\boldsymbol{K}}} \boldsymbol{F}_{\{\boldsymbol{topic\_deviation}\}(\boldsymbol{\Psi^*},\boldsymbol{\Psi})} \quad (\mathbf{19})$$

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

$$\boldsymbol{F}_{\{\boldsymbol{topic\_deviation}\}}(\boldsymbol{\Psi^*}, \boldsymbol{\Psi}) = \frac{\mathbf{1}}{\boldsymbol{K}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{K}} \frac{\mathbf{1}}{|\boldsymbol{\Psi_i}|} \sum_{\boldsymbol{r=1}}^{|\boldsymbol{\Psi_i}|} \left|\left|\boldsymbol{\Psi_i^r} - \boldsymbol{\Psi_i^*}\right|\right|^{\mathbf{2}} \quad (\mathbf{20})$$

where Ψ_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 n-dimensional simplex that is the unit basis vectors

of $\boldsymbol{\mathcal{P}}$. The second group of solutions is generated by combining pairs of corner-located solutions. Letting C_a and C_b 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:

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

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. The linear constraints that must be satisfied 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 point 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. The linear constraints that must be satisfied, 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 points within each cluster center and subsequently performing a random permutation of the selected cluster-center indices. This procedure guarantees that the mutant offspring will also lie within the *T*-dimensional standard simplex.

5.3.7 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:

$$\boldsymbol{\Psi_i} = \{\boldsymbol{\Psi_i^1}, \cdots, \boldsymbol{\Psi_i^{n_i}}\} \quad (\mathbf{22})$$

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

$$\mathbf{S_i} = \{\mathbf{S_i^1}, \cdots, \mathbf{S_i^{n_i}}\} \quad (\mathbf{23})$$

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

$$\boldsymbol{S_i^*} = \frac{\mathbf{1}}{\boldsymbol{n_i}} \sum_{\boldsymbol{r=1}}^{\boldsymbol{n_i}} \boldsymbol{S_{i^r}} \quad (\mathbf{24})$$

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:

$$\mathbf{T_i^*} = \mathbf{arg}\ \max_{\boldsymbol{t} \in \boldsymbol{T}} \boldsymbol{\Psi_{i^*}}(\boldsymbol{t}) \quad (\mathbf{25})$$

5.3.8 Step 8: Output IT Metrics & Consumer Perceptions

Our model generates prevailing topics and clusters in the given corpus, while also producing four key output metrics, namely:

- **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.

5.4 Summary

This chapter aimed at building on the findings presented in the previous chapters and introducing a refined model that mines consumer perceptions from social media as a part of a brand equity assessment exercise. The suggested model was presented through a design science research approach, falling under the improvement DSR quadrant. The model consists of 8 steps that lay out a detailed approach on how to derive specific insights regarding two dimensions of customer based brand equity, namely brand awareness and brand meaning. The refined model is also enhanced from a technical perspective by the introduction of a novel GA for improved data clustering. In the next chapter we empirically apply our model to guide the analysis of data extracted from Twitter and unveil consumer perceptions relating to the UBER transportation network.

Chapter 6

Case Study II: Revealing Consumer Brand Perceptions from Twitter: The case of UBER

Whenever an innovative and disruptive force enters an existing market, usual processes are bound to be reconsidered and change. That certainly seems to be the case with UBER, as recently researchers are trying to figure how much of an impact this new offering is posing on traditional transportation methods.

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 any smartphone user using the app, to search for and request a trip pickup from his exact geographical location, from a designated UBER driver who uses his own vehicle. As of 28 May, 2015, the service was available in 58 countries and 300 cities worldwide[17].

While UBER is transforming the transportation market through the use of innovative technology, it is also gathering vast amounts of data regarding consumer

[17] http://www.uber.com (Last Accessed 26/10/2015)

attitudes and perceptions. One of the official channels for communication with UBER for customer care issues, 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 have to do with, but are certainly not limited to, pricing issues, service complaints, reporting of dangerous driving, user privacy and safety issues. A natural approach towards finding information in this vast pool of data is to filter through keywords search. We instead apply our model to unveil critical information, that of topic clustering, which rather than following a bottom up approach, uses a top-down approach to unveil topics and sentiment towards the brand's action that in turn could be investigated in subtopics and so on.

6.1 Study Background

To illustrate the validity of our model, we unveil consumer perceptions relating to the UBER transportation network. We collected and analyzed a set of over 280.000 tweets, during a three-month period, between January and April 2015, by utilizing the Streaming API of Twitter. This rapid expansion of this revolutionary transport method has not only changed the way people travel in urban environments but also changed the means through which people choose to express their thoughts, complains and general enquiries towards the company. On a 24/7 basis, UBER is the subject of on-going requests through their twitter account, that usually have to do with, but are certainly not limited to, pricing issues, service complaints, reporting of dangerous driving, user privacy and safety issues. The data collection process was focused on gathering tweets that were explicitly referring to the UBER transportation network by performing hash-tag and mentions filtering on the terms "#UBER" and "@UBER". The original collection of tweets was subsequently submitted to a series of data clearing and pre-processing operations. The obtained results provide significant CBBE insights in regards to prevailing factors that influence UBER's users towards the brand.

6.2 Scope & Objectives

Researchers from Marketing have recently stressed the need to deliver actionable insights from social media sources. This new information source is being used to better understand customers, improve future operations and provide metrics that can assist CBBE measurement exercises. This allows the brand to gain the advantage of a holistic and near real time view of their customers, complementing current approaches that heavily rely on traditional data collection and analysis methods such as questionnaires, face to face or telephone interviews, which have a significant time lag.

This shift has resulted in a new pool of data that can be mined in order to understand customer perceptions of the brand. Although positive attributes of customer perceptions can be extracted by various indicators, including current practices, negative attributes can primarily be extracted from consumer perceptions generated from social media due to the nature of the medium that embraces direct communication.

UBER is a perfect example of this paradigm shift as it revolutionizes the way a customer interacts with the brand. UBER's business model relies on the notion that no human interaction is needed during the service (order, ride, pay). The novelty lies on the premise that UBER doesn't need to be present in the physical transaction of the service and it explicitly chooses to direct communication with users through is it's official mobile app and twitter account (@UBER)

This is aligned with current shifts in consumer to company interaction in the digital era. More and more people use smartphones and apps to perform daily activities and communication with the brand shifts towards social – mobile channels. It is evident that more than ever there is now a need for models that utilize big data techniques through mining and analyzing brand-related information from online social networks.

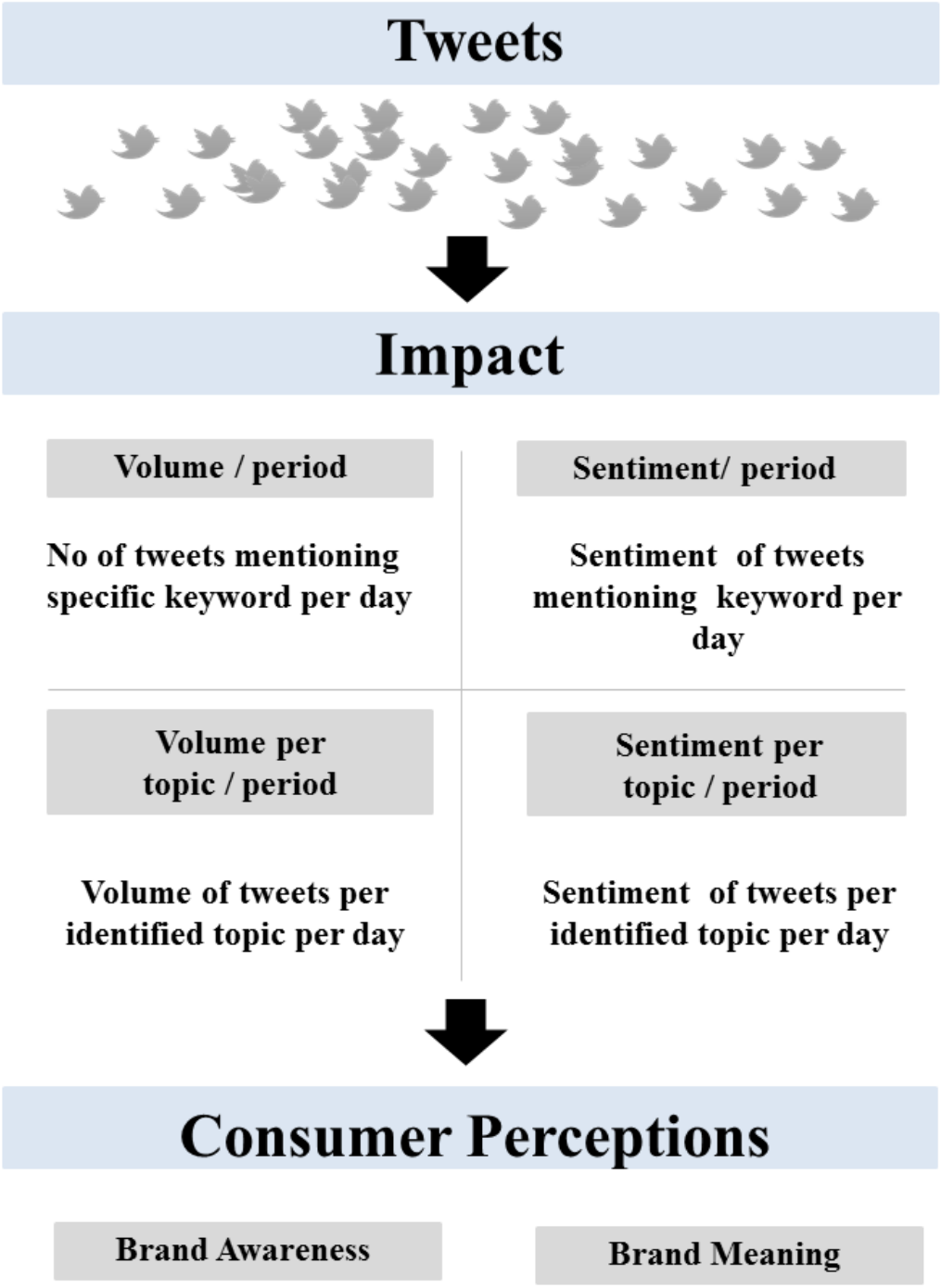


Figure 6.1: Graphical Description of Case Study Objective

6.3 Case Study Steps

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 hash-tags and mentions 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. The data collection process was focused on gathering tweets that were explicitly referring to the UBER transportation network by performing hash-tags and mentions filtering on the terms "#UBER" and "@UBER". 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 and the construction of our corpus as a collection of distinct author documents where each document contained the text from a single tweet. The final version our 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 2 characters. Our final version of our corpus involved 221.958 tweets. Figure 6.2, in particular, depicts the evolution of the daily volume of tweets gathered.

Step 3: VSM-based corpus vectorization: Each document of our 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 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 (see previous chapter), where the parameter γ 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 fist 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 can be found in Appendix 2. 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 6.3 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 6.1.

Step 6: GA Clustering: At this point all tweets are represented as probabilistic mixtures of the *10* prevailing topics in the given corpus and are assigned with 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.

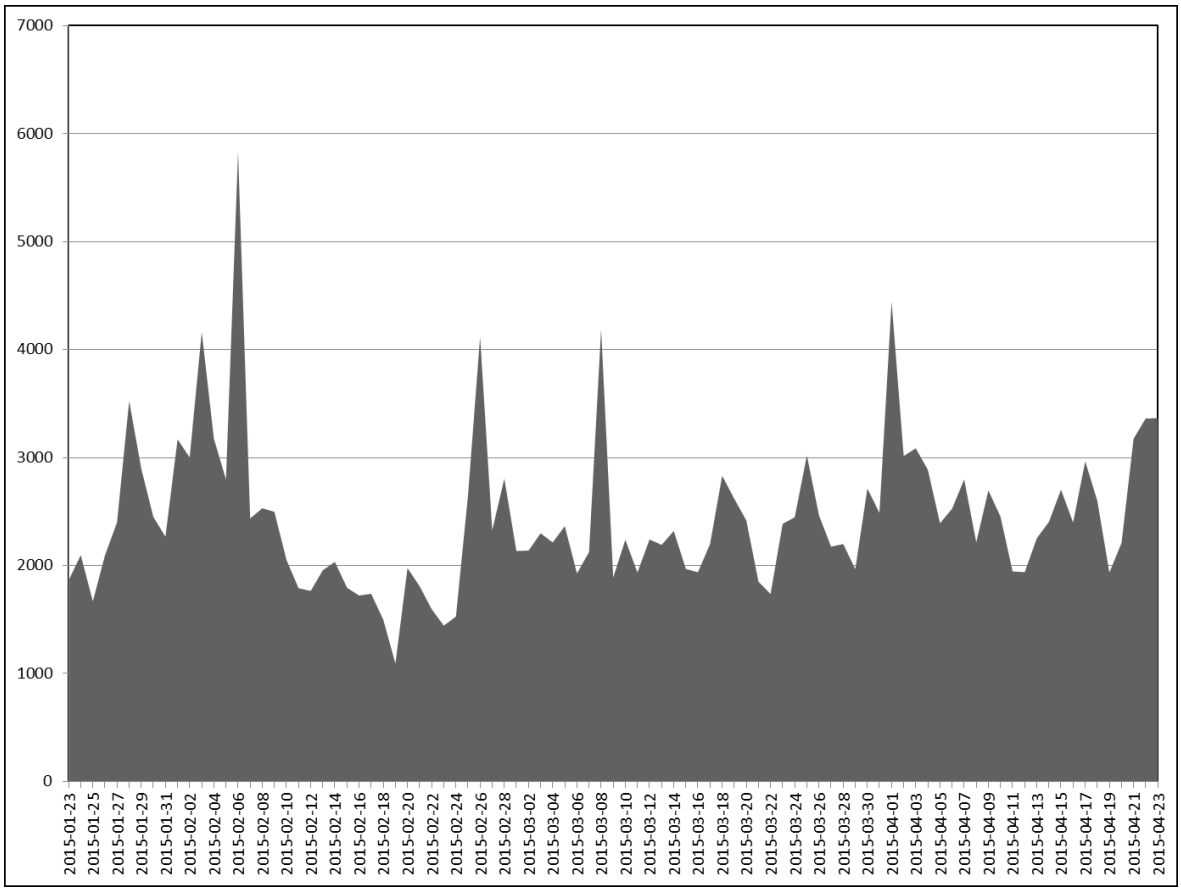


Figure 6.2: Daily Volume of Tweets

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 6.1: Top-10 discussed topics

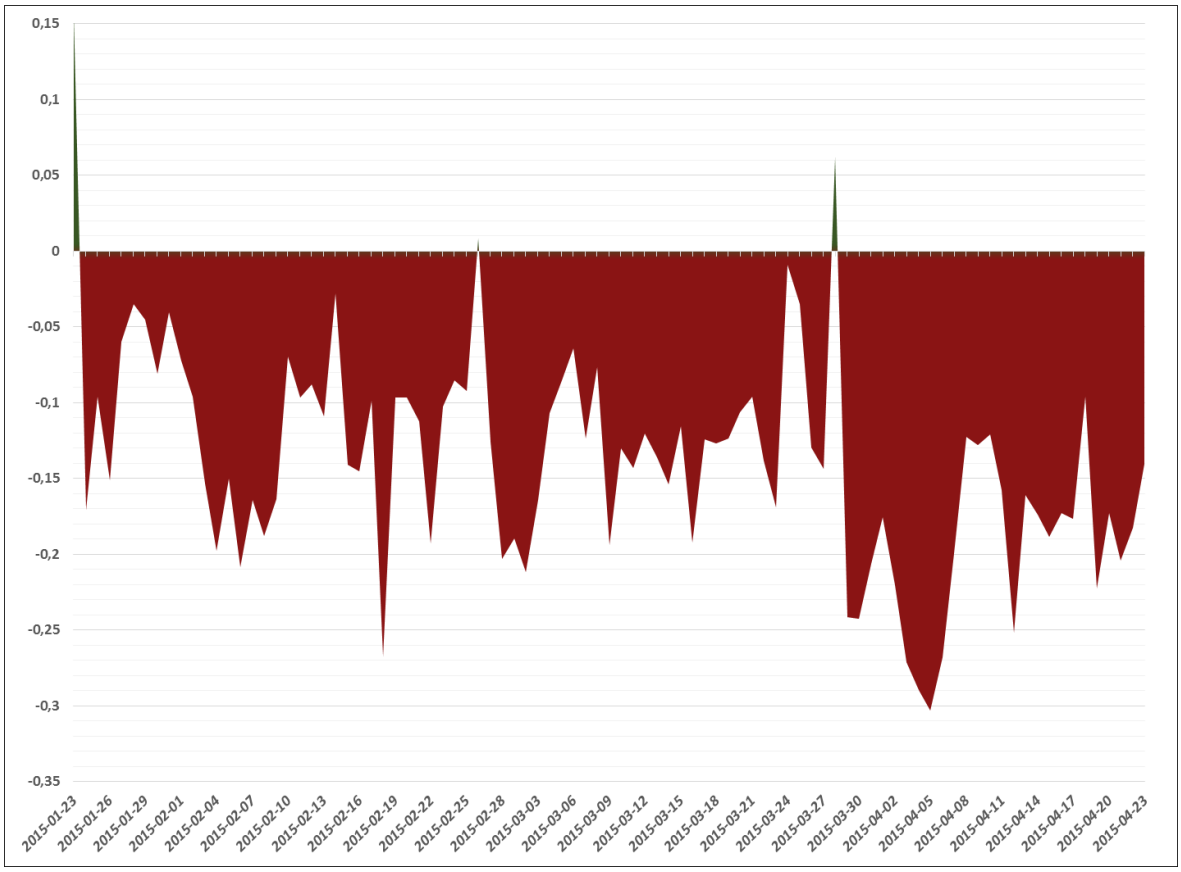


Figure 6.3: Daily Sentiment of Tweets

Results of the GA clustering topic deviation in comparison to K-means are depicted in Table 6.2 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 10 clusters provides significant justification towards inferring that the true number of clusters for this particular dataset is also 10. This, in turn, is no coincidence since **T**=10 is the dimensionality of the LDA-induced semantic space that underlies the given corpus.

The GA – based 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 6.3.

Step 7: Sentiment per cluster: The final step involves sentiment classification of tweets participating in each cluster, depicted at a daily average as per Figure 6.3.

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 6.2: Topic Deviation of GA vs. k-means clustering

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 6.3: Clusters – Prevailing topic of each cluster – No of Tweets

6.4 Results & Discussion

Step 8: Output IT Metrics & Consumer Perceptions: Application of our computational model in this vast pool of "UBER" related data has generated significant insights, which are of particular interest when assessing CBBE for the

given brand. Starting from a collection of more than 280.000 tweets the model manages to generate four key insights with regards to daily sentiment towards the brand, prevailing topics discussed by twitter users, optimal clustering of tweets in semantically coherent clusters under a distinct topic, and an overall average sentiment assessment of each topic per day.

Figure 6.3 aptly depicts the average daily sentiment of users towards "UBER". 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 doesn't 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 6.1 summarizes this information for the given time period. Our data analysis reveals that users pay particular interest, when choosing to express their opinion via twitter, to the following subjects:

1. **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 regards to drivers, dangerous driving, dirty fleet, and incidents of overcharging.
2. **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"
3. **Coupons:** High number of tweets promoting coupons for reduced fares. Usually re-tweeted by an increased number of accounts.
4. **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.
5. **Free codes:** Users' expressing their gratitude for free codes that resulted in free rides.

6. **Support – Help:** Users using twitter as the primary means of contact for support and help issues related to theft, sexual abuse, general complaints and cancellations of service.
7. **Selfies in back of UBER cars:** Users posting selfies with specific hash tags in the back of UBER cars and celebrities which were retweeted posting their selfies.
8. **Surge Pricing:** Huge interest and discussion about UBER's surge pricing algorithm which in many cases frustrates users due to overcharging.
9. **Women & UBER:** Specific discussion sparked by UBER's initiative to create jobs for women.
10. **#UBERed:** Trending hash tag applied when having a bad experience or worse on UBER. This type of expression has become a meme between UBER users especially when resulting to 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. This metric indicates user's 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 (Figure 6.4 and Figure 6.5).

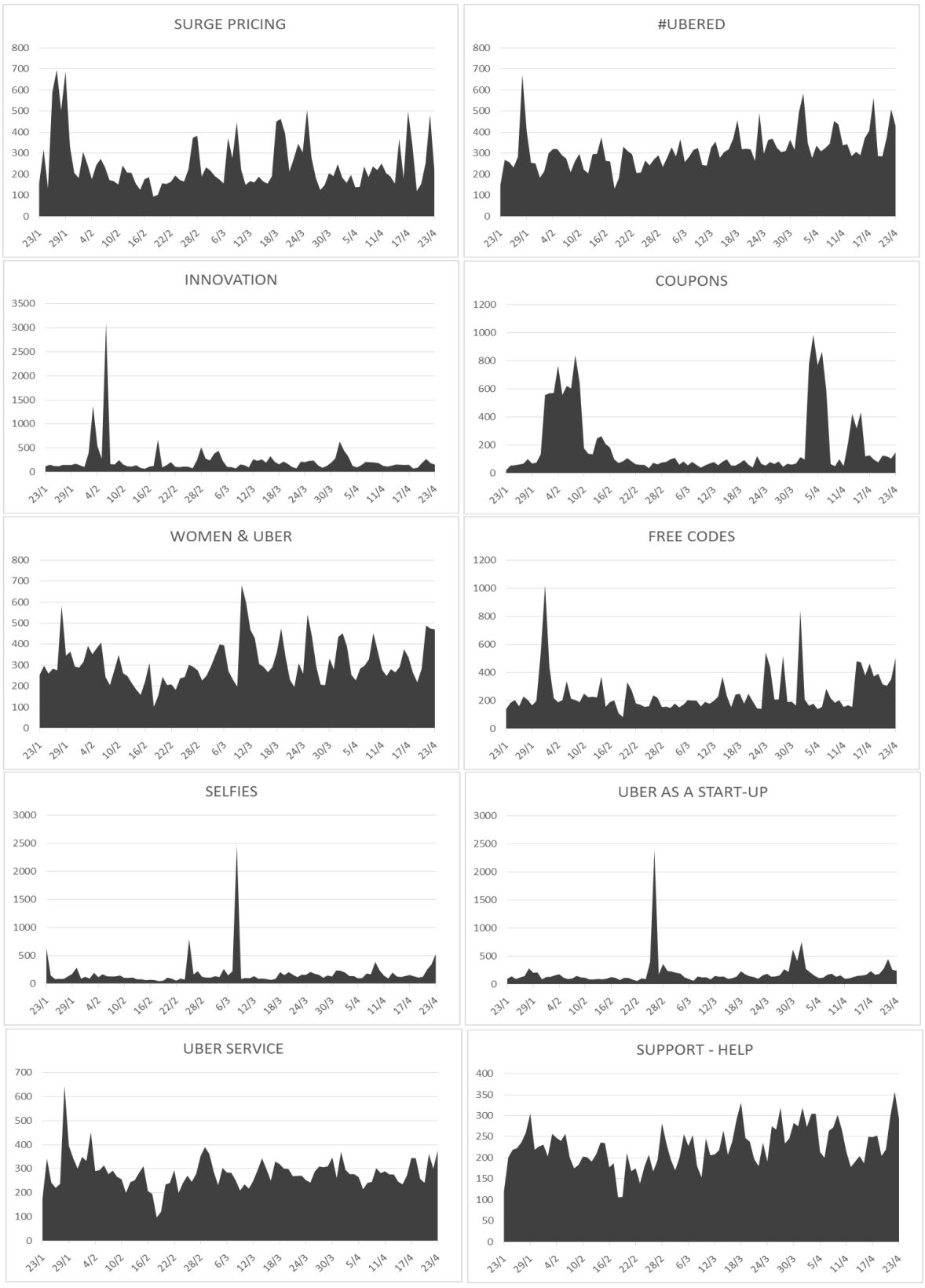


Figure 6.4: Daily Volume of Tweets per Topic

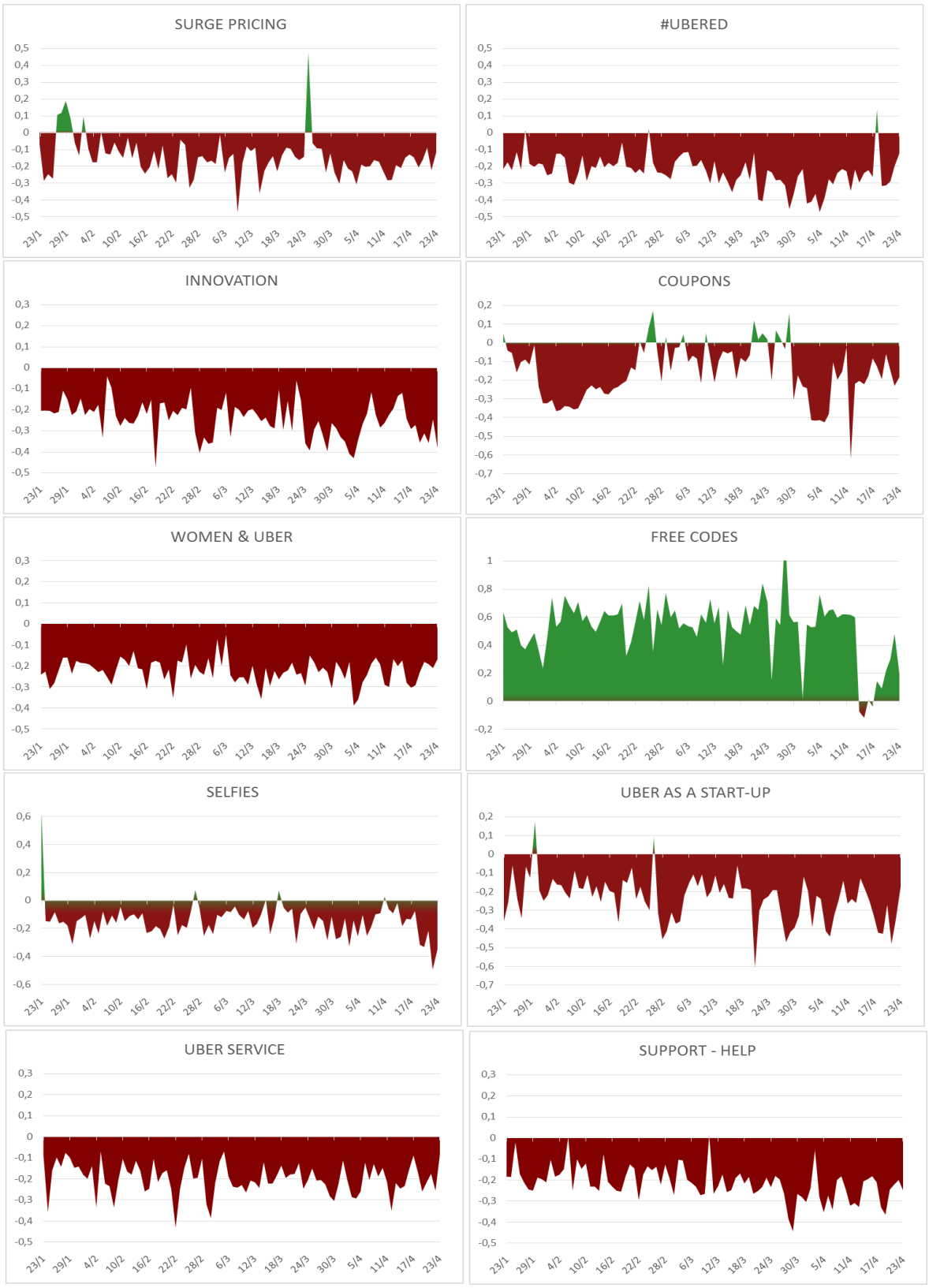


Figure 6.5: Daily Sentiment of Tweets per Topic

6.5 Summary

In this chapter, we empirically applied a computational model to guide the analysis of over 280,000 tweets in regards to the UBER transportation network and the users who included the specific brand as part of their tweeting activity. In particular, we evaluated the model with the addition of a novel GA which significantly improves clustering of tweets in terms of semantic coherence as measured by the utilized topic deviation metric when compared against the traditional k-means clustering algorithm for a series of different experimentation scenarios with an increasing number of clusters to be found. Our results indicate that the dataset is inherently organized in 10 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 drawing upon Big Data and Machine Learning techniques for improved data analysis. 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. From a Marketing perspective we stress the need for brands to complement existing CBBE assessment exercises with insights generated from mining consumer perceptions from SMN.

This study also presents some insightful managerial implications. Marketers 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 results 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.

Chapter 7

Conclusions, Limitations & Future Research

The final chapter of this thesis serves as a concluding note to the research presented in the previous chapters. The first section provides a summary of the work discussed in this dissertation. We continue with the theoretical and managerial contributions and discuss limitations of the proposed approach. Finally, we conclude by discussing future research directions and pathways, which may help researchers, extend and push the presented work further.

7.1 Introduction

This thesis studied research techniques from Marketing, Information Systems and Machine Learning and examined how these disciplines could be intertwined to solve the problem of generating customer based brand equity insights through social media. The starting point has been to carefully study the marketing construct of brand equity, the dimensions that govern its nature and the various approaches introduced in literature for measuring its impact.

A careful look at the literature identifies the growing importance of brand equity as a marketing construct and the ongoing attempts by researchers to universally define the term and provide methodologies for measuring its different dimensions. A

common characteristic of the introduced approaches, which are still used today by marketers, is that they focus on the multidimensionality of the construct, collect data through traditional methods and use confirmatory factor analysis with structural equations modeling for evaluating their models.

At the time this research set off, marketing practitioners were struggling to cope with the proliferation of a new emerging medium, broadly defined by researchers in Information Systems as Social Media Networks. Researchers such as Bruhn, Schoenmueller & Schafer (2012) were among the first to question the relative impact of brand communication on brand equity through traditional media as compared to social media. In parallel researchers from Computer Science such as Melville, Sindhwani, and Lawrence (2009) highlighted the need to harness the power of social media, in particular the opportunity to apply machine learning and data techniques, to infer marketing insights from such mediums (web blogs at the time).

Motivated by this disruption in marketing, this research initially set out to examine approaches for mining and analyzing data from online social networks such as Facebook and Twitter. Our initial intention was to develop a framework that would allow researchers to test hypotheses or validate new theories by exploiting the enormous amounts of real-time data-driven information that combined topic and sentiment detection approaches, aiming to elicit insights that govern the generation process of consumer satisfaction. As such, a proposed method that relied on a set of state-of-the-art machine learning techniques such as, data collection, topic modeling and sentiment classification was introduced aiming to probe for discussion themes that were highly influential to the formulation of positive and negative opinions.

The proposed model was subsequently tested through the application of a real life case study, mining Twitter for customer satisfaction rates, in two of the leading Telecommunication providers in the US, AT&T and Verizon. In particular its main goal was to examine if mobile wireless carriers could benefit from performing topic modeling and sentiment analysis through social media networks in order to gain

insights about the performance of their service. Although the results obtained were of particular interest, they sparked a series of additional questions that needed to be addressed, in regards to the model's accuracy and validity. In particular, the findings indicated the need for the model to further focus on a specific marketing construct, enhance the computational engine and perform a new series of experiments to test its validity.

Chapter 5 presented the refined model which relied on a design science research framework to construct the proposed informational technology artifact that utilized modern machine learning techniques to build a model of data processing, topic modeling and sentiment classification. In particular the model introduced, consists of 8 steps that lay out a detailed approach on how to derive specific insights, from mining consumer brand perceptions from SMN, regarding two dimensions of customer based brand equity, brand awareness and brand meaning. The efficiency and validity of the proposed computational model was tested through its application in a second case study, examining consumer perceptions for the UBER brand.

The next section opens a discussion on the main theoretical and managerial contributions of the research presented in this thesis.

7.2 Contribution

The research presented in this thesis initially set out to delve in current marketing practices of assessing customer based brand equity and identify the technological advancements in machine learning that could contribute to the paradigm shift in differentiated assessment approaches. Given its interdisciplinary nature, the research was carefully grounded in Marketing and Information Systems literature, while also drawing on state of the art Machine Learning techniques.

Our research inquiry differentiated from current marketing research approaches as its main goal was to develop a computational model for mining consumer

perceptions from social media channels that could identify two important customer based brand equity dimensions. In doing so, our research followed a construct development methodology, introduced in the form of a novel IT artifact, following a design science research approach. The proposed approach varies significantly from current practices that attempt to address the same problem through qualitative research approaches that fail to harness state of the art big data and machine learning techniques for analyzing data in near real time. Hence, one contribution of this research is being among the first to develop a computational model for assessing a specific marketing construct, such as customer based brand equity, by collecting and analyzing data from social media networks.

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 Marketing perspective, we contend that traditional methods of data gathering provide a static and sometimes skewed measure of brand equity and stress the need to combine a business-driven approach to structuring an SMA project with detailed, state-of-the-art, actionable methods for performing analytics in specific SMN datasets. As such we stress the need for brands to complement existing CBBE assessment exercises with insights generated from mining consumer perceptions from SMN as introduced in our computational model.

The results obtained, through the application of our model in two distinct real life case studies, confirmed our research goal and comply with the research objective. For example, the case study presented in chapter 4 revealed that the latest advertisement spot by AT&T contributed to breeding positive social sentiment values for the firm's customers. Subsequently the second case study, presented in chapter 6, indicated that the dataset was inherently organized in 10 semantically focused clusters, each one minimally distributed around a unique topic. A direct consequence of this fact was that we could associate each cluster and corresponding

topic with an average sentiment value, thus unraveling the public attitude against particular aspects of the brand under investigation.

From a Machine Learning perspective, we addressed the problem of topic clustering, through the utilization of a novel genetic algorithm approach which is highly scalable on large volumes of textual data, by introducing a centroid-based encoding scheme. The novelty of our proposed genetic algorithm approach lies primarily upon the adaptation of the centroid-based encoding scheme, in the sense that cluster assignments are implicitly extracted by assigning each data point to the nearest cluster center. Results from the case study in chapter 6, 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 (Table 6.2).

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 in chapter 6 revealed that twitter users are particularly negative towards Uber's service, support, and strategy of expansion in the market. On the contrary, users 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.

In sum, the theoretical and managerial contributions of this research outline the need for marketing researchers and practitioners to adopt state of the art machine learning techniques and methods, when faced to assess specific marketing constructs, such as customer based brand equity. Our research contributes towards this end by outlining a step by step computational model that starts from a marketing perspective, describes the necessary technological steps and concludes with the insights that can be generated. As such, this study paves the way for future

researchers to build on this model and investigate similar techniques that either assess the construct of CBBE more efficiently or assess other marketing constructs with similar success.

7.3 Limitations

Our study has several limitations that can also act as incitement for future research. In this section we discuss the limitations that percolate this research and ground our discussion for the next section, where we explore how these limitations could be overcome by future research endeavors.

The first limitation deals with the fact that both case studies used Twitter as the primary source for collecting data from users. This in itself generates various implications as different social media platforms have their own technical and business requirements for collecting, analyzing and interpreting data generated from their user base. A major limitation faced in both case studies was the inability to use Twitter's Search API for collecting historical data. Access to the specific API is granted to a limited number of research organizations worldwide, which we were unable to be granted, even though we inquired in several occasions. We thus had to collect tweets through the use of Twitter Streaming API, which collects real-time generated data on a 24/7 basis. Other types of social media platforms such as Facebook, YouTube, Instagram, etc. also provide their own APIs for data collection, each one with its own technical limitations. Our proposed model is designed in such a way that can be applicable to any medium that provides the relevant API for data extraction, after the researcher consults the relevant technical guidelines for the medium he wishes to collect data from.

A second limitation regards alternate aspects of tweeting activity such as hearts (previously known as favorites), re-tweets without comment and meta-attributes such as number of followers, geo-location, etc. These meta-attributes are an additional pool of information that could further enhance similar models, which for

our purposes where not included. Our model explicitly focuses on the semantic attributes of the tweets (text analysis) which capture comments and RTs 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 third limitation is that results derived from mining any online medium for sentiment and scope, don't necessarily align with real world viewpoints, mainly due to the characteristics of the population choosing to express views from the specific mean. Population characteristics of social media users indicate usage mainly from younger audiences, living in areas of high income and advanced technological and network infrastructure. A significant proportion of Social Media users are also in forms of bots, aggregators or noise creators. As a direct consequence, a percentage of collected tweets that mention the brand under investigation have no significance in the analysis (varying based on the query).

A major limitation regards the behavioral activity of users choosing to express their experience with the brand. 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. As various research studies indicate (Hutter et al. 2013; McCarthy et al. 2014; Hudson et al. 2015), users choosing to express their experience through the web are most likely to do so after a negative experience. A consequential effect is the ongoing challenge from a technical perspective to detect irony in sentiment analysis exercises. Although various models have and continue to be presented (Bosco, Patti, and Bolioli 2013; Montoyo, MartíNez-Barco, and Balahur 2012; Ghosh et al. 2015; Wallace 2015) in the Computing discipline, it still is a major challenge to detect ironic activity in text, especially using expressions and memes as a direct result of social media communication.

Finally, two limitations governing the computational engine of our model have to do with training of the algorithm and determining the optimal number of topics and clusters. Every machine learning algorithm is essentially as good as its training data. For the first case study we employed eight undergraduate students to manually grade data in terms of their sentiment. Nationality and base of residence of the students (Greece) ensured bias towards a given company, as all eight students had never used any of the company's services. Nevertheless, our results showed that training of the algorithm could have performed better if the manual grading was done by native speakers and in a more balanced way for the two brands under investigation (AT&T, Verizon). We followed this approach in the second case study but again a larger amount of scored tweets would have resulted in even better results. Finally, 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.

We urge future researches wishing to apply and extend our proposed model, to take the limitations discussed above into account in the experimental applications they chose to proceed. To aid them in their future endeavors we discuss some areas of interest in the section that follows.

7.4 Future Research Directions

> *"Big Data is the biggest game-changing opportunity for marketing and sales since the Internet went mainstream almost 20 years ago"* [18]

That statement featured in a study by McKinsey portrays the scene under which marketing researchers and practitioners operate these days. It has also enabled

[18] [Online]: http://www.forbes.com/sites/mckinsey/2013/07/22/big-data-analytics-and-the-future-of-marketing-sales/#2a1cc291344d [Last Accessed:] 11-August-2016

rigorous research in the fields of digital marketing and big data analytics to emerge in recent years. Our proposed model supports the ever expanding chain of research and may spark future research in the field of Social Media Analytics. Future directions are listed but certainly not limited, in the following paragraphs.

The proliferation of social media is still ongoing. Traditional applications such as Facebook, Twitter and LinkedIn are now facing competition from new and emerging apps such as Instagram, SnapChat, Tumblr and social messaging apps such as WhattsApp, Viber and WeChat. We urge future researchers to consider applying our model to media as such and take advantage of the different properties that each application provides (geo-location, hearts, likes, emoticons) for further analysis of non-semantic properties, in regards to different CBBE dimensions.

From a marketing perspective, it would be interesting to consider applying the model in different market verticals that would help generate CBBE insights in various fields of business. Furthermore it would also be interesting to consider assessing a product rather than a service brand and possibly tweak the CBBE dimensions assessed, shifting away from Berry's model to other models introduced in literature. This would provide the opportunity to extend the claim that traditional CBBE assessment approaches should be complemented with models that harness big data from social media. A means towards that end could also be a study where results of our model are compared against results from traditional methods for the same brand.

A third direction of future research that would be stimulating would be to expand the computational engine of the model to infer insights from network topological properties. Algorithms in the field of Network Science allow us to reveal valuable insights from dynamic systems yet still few studies utilize the semantic information residing in social media networks to present consumer insights through time. Semantic information residing in social media is a critical factor in regards to detection of communities in complex systems which may provide additional CBBE insights.

Finally, drawing from our findings from the computational modules, the model could be extended by focusing on incorporating the temporal parameter within the clustering process in order to develop a group monitoring mechanism. This could be achieved by further experimenting with different values for the initial topic numbers in regards to the optimal clustering set. In parallel future researchers could experiment by further expanding the modules for better sentiment and topic clustering results.

References

Aaker, David A. 1996. “Measuring Brand Equity Across Products and Markets.” *California Management Review* 38 (3): 102–20.

Aaker, David A. 1992. “The Value of Brand Equity.” *Journal of Business Strategy* 13 (4): 27–32.

Aggarwal, Charu C., and ChengXiang Zhai. 2012. “A Survey of Text Clustering Algorithms.” In *Mining Text Data*, edited by Charu C. Aggarwal and ChengXiang Zhai, 77–128. Boston, MA: Springer US.

Agustın-Blas, L.E., S. Salcedo-Sanz, S. Jiménez-Fernández, L. Carro-Calvo, J. Del Ser, and J.A. Portilla-Figueras. 2012. “A New Grouping Genetic Algorithm for Clustering Problems.” *Expert Systems with Applications* 39 (10): 9695–9703.

Alan, R. Hevner von, Salvatore T. March, Jinsoo Park, and Sudha Ram. 2004. “Design Science in Information Systems Research.” *MIS Quarterly* 28 (1): 75–105.

Amblee, Naveen, and Tung Bui. 2011. “Harnessing the Influence of Social Proof in Online Shopping: The Effect of Electronic Word of Mouth on Sales of Digital Microproducts.” *International Journal of Electronic Commerce* 16 (2): 91–114.

Aral, Sinan, and Dylan Walker. 2011. “Creating Social Contagion through Viral Product Design: A Randomized Trial of Peer Influence in Networks.” *Management Science* 57 (9): 1623–39.

Asur, Sitaram, and Bernardo A. Huberman. 2010. “Predicting the Future with Social Media.” In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, 1:492–99. IEEE.

Bandyopadhyay, Sanghamitra, and Sankar Kumar Pal. 2007. *Classification and Learning Using Genetic Algorithms: Applications in Bioinformatics and Web Intelligence*. Springer Science & Business Media.

Berry, David M. 2011. “The Computational Turn: Thinking about the Digital Humanities.” *Culture Machine* 12 (0): 2.

Berry, Leonard L. 2000. “Cultivating Service Brand Equity.” *Journal of the Academy of Marketing Science* 28 (1): 128–37.

Beverland, Michael, Adam Lindgreen, Julie Napoli, David Ballantyne, and Robert Aitken. 2007. “Branding in B2B Markets: Insights from the Service-Dominant Logic of Marketing.” *Journal of Business & Industrial Marketing* 22 (6): 363–71.

Bhattacherjee, Anol. 2012. "Social Science Research: Principles, Methods, and Practices." http://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=1002&context=oa_textbooks.

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55 (4): 77–84.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *The Journal of Machine Learning Research* 3: 993–1022.

Bollen, Johan, Huina Mao, and Xiaojun Zeng. 2011. "Twitter Mood Predicts the Stock Market." *Journal of Computational Science* 2 (1): 1–8.

Bosco, Cristina, Viviana Patti, and Andrea Bolioli. 2013. "Developing Corpora for Sentiment Analysis: The Case of Irony and Senti-Tut." *IEEE Intelligent Systems* 28 (2): 55–63.

boyd, danah, and Kate Crawford. 2012. "Critical Questions for Big Data." *Information, Communication & Society* 15 (5): 662–79. doi:10.1080/1369118X.2012.678878.

Bruhn, Manfred, Verena Schoenmueller, and Daniela B. Schäfer. 2012. "Are Social Media Replacing Traditional Media in Terms of Brand Equity Creation?" *Management Research Review* 35 (9): 770–90.

Buchanan, Elizabeth. 2012. "Ethical Decision-Making and Internet Research." http://www.dphu.org/uploads/attachements/books/books_5612_0.pdf.

Cai, Deng, Qiaozhu Mei, Jiawei Han, and Chengxiang Zhai. 2008. "Modeling Hidden Topics on Document Manifold." In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, 911–20. ACM.

Callarisa, Luis, Javier Sánchez García, John Cardiff, and Alexandra Roshchina. 2012. "Harnessing Social Media Platforms to Measure Customer-Based Hotel Brand Equity." *Tourism Management Perspectives* 4 (October): 73–79.

Chang, Dong-Xia, Xian-Da Zhang, and Chang-Wen Zheng. 2009. "A Genetic Algorithm with Gene Rearrangement for K-Means Clustering." *Pattern Recognition* 42 (7): 1210–22.

Chernatony, Leslie De, Fiona Harris, and George Christodoulides. 2004. "Developing a Brand Performance Measure for Financial Services Brands." *The Service Industries Journal* 24 (2): 15–33.

Chernatony, Leslie De, Malcolm McDonald, and Malcolm McDonald. 1992. *Creating Powerful Brands: The Strategic Route to Success in Consumer, Industrial and Service Markets*. Butterworth-Heinemann Oxford.

Christodoulides, George, and Leslie De Chernatony. 2010. "Consumer-Based Brand Equity Conceptualization and Measurement: A Literature Review." *International Journal of Research in Marketing* 52 (1): 43–66.

Christodoulides, George, Leslie De Chernatony, Olivier Furrer, Eric Shiu, and Temi Abimbola. 2006. "Conceptualising and Measuring the Equity of Online Brands." *Journal of Marketing Management* 22 (7-8): 799–825.

Chui, Michael, James Manyika, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Hugo Sarrazin, Geoffrey Sands, and Magdalena Westergren. 2012. "The Social Economy: Unlocking Value and Productivity through Social Technologies." *McKinsey Global Institute* 4.

Clark, Eleanor, and Kenji Araki. 2011. "Text Normalization in Social Media: Progress, Problems and Applications for a Pre-Processing System of Casual English." *Procedia-Social and Behavioral Sciences* 27: 2–11.

Clifton, C., R. Cooley, and J. Rennie. 2004. "TopCat: Data Mining for Topic Identification in a Text Corpus." *IEEE Transactions on Knowledge and Data Engineering* 16 (8): 949–64.

Cobb-Walgren, Cathy J., Cynthia A. Ruble, and Naveen Donthu. 1995. "Brand Equity, Brand Preference, and Purchase Intent." *Journal of Advertising* 24 (3): 25–40.

Cortes, Corinna, and Vladimir Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20 (3): 273–97.

Coulter, Keith S, Manfred Bruhn, Verena Schoenmueller, and Daniela B Schäfer. 2012. "Are Social Media Replacing Traditional Media in Terms of Brand Equity Creation?" *Management Research Review* 35 (9): 770–90.

Culotta, Aron, and Jennifer Cutler. 2016. "Mining Brand Perceptions from Twitter Social Networks." *Marketing Science* 35 (3): 343–62.

Davidson, Alistair, and Jonathan Copulsky. 2006. "Managing Webmavens: Relationships with Sophisticated Customers via the Internet Can Transform Marketing and Speed Innovation." *Strategy & Leadership* 34 (3): 14–22.

Davis, Donna F., Susan L. Golicic, and Adam J. Marquardt. 2008. "Branding a B2B Service: Does a Brand Differentiate a Logistics Service Provider?" *Industrial Marketing Management* 37 (2): 218–27.

Davis, Lawrence. 1991. *Handbook of Genetic Algorithms*. Van Nostrand Reinhold.

Deerwester, Scott C., Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. "Indexing by Latent Semantic Analysis." *JAsIs* 41 (6): 391–407.

Dellarocas, Chrysanthos, and Charles A. Wood. 2008. "The Sound of Silence in Online Feedback: Estimating Trading Risks in the Presence of Reporting Bias." *Management Science* 54 (3): 460–76.

Duan, Changming, and Clara E. Hill. 1996. "The Current State of Empathy Research." *Journal of Counseling Psychology* 43 (3): 261.

Fan, Weiguo, and Michael D. Gordon. 2014. "The Power of Social Media Analytics." *Communications of the ACM* 57 (6): 74–81.

Farquhar, P. H. 1990. "'Managing Brand equity'Journal of Advertising Research, Vol. 30, August-September, Pp." *RC7-RC12*.

Farquhar, Peter H. 1989. "Managing Brand Equity." *Marketing Research* 1 (3).

Fersini, E, E Messina, and FA Pozzi. 2014. "Sentiment Analysis: Bayesian Ensemble Learning." *Decision Support Systems* 68: 26–38.

French, Alan, and Gareth Smith. 2013. "Measuring Brand Association Strength: A Consumer Based Brand Equity Approach." *European Journal of Marketing* 47 (8): 1356–67.

Gallaugher, John, and Sam Ransbotham. 2010. "Social Media and Customer Dialog Management at Starbucks." *MIS Quarterly Executive* 9 (4).

Galliers, Robert. 1992. *Information Systems Research: Issues, Methods and Practical Guidelines*. Blackwell Scientific.

Garai, Gautam, and B.B Chaudhuri. 2004. "A Novel Genetic Algorithm for Automatic Clustering." *Pattern Recognition Letters* 25 (2): 173–87.

Ghosh, Aniruddha, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. "Semeval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter." In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 470–78.

Goldberg, David E. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*. 1st ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Goldberg, David E., and John H. Holland. 1988. "Genetic Algorithms and Machine Learning." *Machine Learning* 3 (2): 95–99.

Green, Paul E, Ronald E Frank, and Patrick J Robinson. 1967. "Cluster Analysis in Test Market Selection." *Management Science* 13 (8): B–387.

Gregor, Shirley, and Alan R Hevner. 2013. "Positioning and Presenting Design Science Research for Maximum Impact." *MIS Quarterly* 37 (2): 337–55.

Gregor, Shirley, and David Jones. 2007. "The Anatomy of a Design Theory." *Journal of the Association for Information Systems* 8 (5).

Gruhl, Daniel, Ramanathan Guha, Ravi Kumar, Jasmine Novak, and Andrew Tomkins. 2005. "The Predictive Power of Online Chatter." In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 78–87. ACM.

Guha, S., A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan. 2003. "Clustering Data Streams: Theory and Practice." *IEEE Transactions on Knowledge and Data Engineering* 15 (3): 515–28.

Hassan Zadeh, Amir, and Ramesh Sharda. 2014. "Modeling Brand Post Popularity Dynamics in Online Social Networks." *Decision Support Systems* 65 (September): 59–68.

He, Hong, and Yonghong Tan. 2012. "A Two-Stage Genetic Algorithm for Automatic Clustering." *Neurocomputing* 81 (April): 49–59.

Hoffman, Donna L., and Marek Fodor. 2010. "Can You Measure the ROI of Your Social Media Marketing." *MIT Sloan Management Review* 52 (1): 41–49.

Hofmann, Thomas. 1999. "Probabilistic Latent Semantic Indexing." In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 50–57. ACM.

Holland, John H. 1973. "Holland, J.: Genetic Algorithms and the Optimal Allocation of Trials. SIAM J. Computing 2, 88-105." *SIAM Journal of Computing* 2 (2): 88–105.

Hudson, Simon, Martin S. Roth, Thomas J. Madden, and Rupert Hudson. 2015. "The Effects of Social Media on Emotions, Brand Relationship Quality, and Word of Mouth: An Empirical Study of Music Festival Attendees." *Tourism Management* 47: 68–76.

Hutter, Katja, Julia Hautz, Severin Dennhardt, and Johann Füller. 2013. "The Impact of User Interactions in Social Media on Brand Awareness and Purchase Intention: The Case of MINI on Facebook." *Journal of Product & Brand Management* 22 (5/6): 342–51.

"IBM Global CMO Study." 2014. CTZ20. February 24. https://www14.software.ibm.com/webapp/iwm/web/signup.do?source=swg-smartercommerce-emm&S_PKG=cs_cmo_study.

"IBM Insights from the IBM Global C-Suite Study." 2014. Ct512. May 7. http://www-935.ibm.com/services/us/en/c-suite/csuitestudy2013/.

Jain, A. K., M. N. Murty, and P. J. Flynn. 1999. "Data Clustering: A Review." *ACM Computing Surveys* 31 (3): 264–323.

Jain, Anil K. 2010. "Data Clustering: 50 Years beyond K-Means." *Pattern Recognition Letters* 31 (8): 651–66.

Jevons, Colin, and Mark Gabbott. 2000. "Trust, Brand Equity and Brand Reality in Internet Business Relationships: An Interdisciplinary Approach." *Journal of Marketing Management* 16 (6): 619–34.

Kane, Gerald C, Maryam Alavi, Giuseppe Labianca, and Stephen P Borgatti. 2014. "What's Different about Social Media Networks? A Framework and Research Agenda." *MIS Quarterly* 38 (1): 275–304.

Kapferer, Jean-Noel. 2012. *The New Strategic Brand Management: Advanced Insights and Strategic Thinking*. Kogan page publishers.

Keller, Kevin Lane. 1993. "Conceptualizing, Measuring, and Managing Customer-Based Brand Equity." *The Journal of Marketing*, 1–22.

Kuhn, Adrian, Stéphane Ducasse, and Tudor Gírba. 2007. "Semantic Clustering: Identifying Topics in Source Code." *Information and Software Technology* 49 (3): 230–43.

Lassar, Walfried, Banwari Mittal, and Arun Sharma. 1995. "Measuring Customer-Based Brand Equity." *Journal of Consumer Marketing* 12 (4): 11–19.

Leuthesser, Lance. 1988. *Defining, Measuring, and Managing Brand Equity: A Conference Summary by*. Marketing Science Institute.

Liang, Ting-Peng, and Efraim Turban. 2011. "Introduction to the Special Issue Social Commerce: A Research Framework for Social Commerce." *International Journal of Electronic Commerce* 16 (2): 5–14.

Lovelock, Christopher H., Jochen Wirtz, and Patricia Chew. 2009. "Essentials of Services Marketing."

Lu, E.H.-C., V.S. Tseng, and P.S. Yu. 2011. "Mining Cluster-Based Temporal Mobile Sequential Patterns in Location-Based Service Environments." *IEEE Transactions on Knowledge and Data Engineering* 23 (6): 914–27.

Luo, Xueming, Jie Zhang, and Wenjing Duan. 2013. "Social Media and Firm Equity Value." *Information Systems Research* 24 (1): 146–63.

M. Mahdavi, M. Haghir Chehreghani. 2008. "Novel Meta-Heuristic Algorithms for Clustering Web Documents." *Applied Mathematics and Computation* 201 (1-2): 441–51.

Mahajan, Vijay, Vithala R. Rao, and Rajendra K. Srivastava. 1990. "Development, Testing, and Validation of Brand Equity under Conditions of Acquisition and Divestment." In *Managing Brand Equity: A Conference Summary Report*, 14–15. Marketing Science Institute.

March, Salvatore T., and Gerald F. Smith. 1995. "Design and Natural Science Research on Information Technology." *Decision Support Systems* 15 (4): 251–66.

Maulik, Ujjwal, and Sanghamitra Bandyopadhyay. 2000. "Genetic Algorithm-Based Clustering Technique." *Pattern Recognition* 33 (9): 1455–65.

McCarthy, Jeff, Jennifer Rowley, Catherine Jane Ashworth, and Elke Pioch. 2014. "Managing Brand Presence through Social Media: The Case of UK Football Clubs." *Internet Research* 24 (2): 181–204.

McDonald, Malcolm HB, Leslie de Chernatony, and Fiona Harris. 2001. "Corporate Marketing and Service Brands-Moving beyond the Fast-Moving Consumer Goods Model." *European Journal of Marketing* 35 (3/4): 335–52.

McGivern, Yvonne. 2009. *The Practice of Market Research: An Introduction*. Pearson Education.

McLaren, Tim S., Milena M. Head, Yufei Yuan, and Yolande E. Chan. 2011. "A Multilevel Model for Measuring Fit Between a Firm's Competitive Strategies and Information Systems Capabilities." *MIS Quarterly* 35 (4).

Mei, Qiaozhu, and ChengXiang Zhai. 2006. "A Mixture Model for Contextual Text Mining." In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 649–55. ACM.

Melville, Prem, Vikas Sindhwani, and R. Lawrence. 2009. "Social Media Analytics: Channeling the Power of the Blogosphere for Marketing Insight." *Proc. of the WIN*.

Mikheev, Andrei. 2000. "Document Centered Approach to Text Normalization." In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 136–43. ACM.

Montoyo, Andrés, Patricio MartíNez-Barco, and Alexandra Balahur. 2012. "Subjectivity and Sentiment Analysis: An Overview of the Current State of the Area and Envisaged Developments." *Decision Support Systems* 53 (4): 675–79.

Moore, Jeri. 1993. "Building Brands across Markets: Cultural Differences in Brand Relationships within the European Community." *Brand Equity & Advertising: Advertising's Role in Building Strong Brands* 1: 31.

Murthy, Chivukula A., and Nirmalya Chowdhury. 1996. "In Search of Optimal Clusters Using Genetic Algorithms." *Pattern Recognition Letters* 17 (8): 825–32.

Onkvisit, Sak, and John J. Shaw. 1989. "Service Marketing: Image, Branding, and Competition." *Business Horizons* 32 (1): 13–18.

"Oscars Senti-Meter." 2014. *Graphics.latimes.com*. Accessed March 31. http://graphics.latimes.com/senti-meter/.

Pacheco, Joaquın A. 2005. "A Scatter Search Approach for the Minimum Sum-of-Squares Clustering Problem." *Computers & Operations Research* 32 (5): 1325–35.

Palka, Wolfgang, Key Pousttchi, and Dietmar G. Wiedemann. 2009. "Mobile Word-of-mouth–A Grounded Theory of Mobile Viral Marketing." *Journal of Information Technology* 24 (2): 172–85.

Pappu, Ravi, Pascale G. Quester, and Ray W. Cooksey. 2005. "Consumer-Based Brand Equity: Improving the Measurement–empirical Evidence." *Journal of Product & Brand Management* 14 (3): 143–54.

Park, Chan Su, and Vinay Srinivasan. 1994. "A Survey-Based Method for Measuring and Understanding Brand Equity and Its Extendibility." *Journal of Marketing Research (JMR)* 31 (2).

Phillips, Estelle, and Derek Salman Pugh. 2005. *How to Get a PhD: A Handbook for Students and Their Supervisors*. Maidenhead: Open University Press.

Pizzuti, Clara. 2012. "A Multiobjective Genetic Algorithm to Find Communities in Complex Networks." *Evolutionary Computation, IEEE Transactions on* 16 (3): 418–30.

Premalatha, K., and A. M. Natarajan. 2010. "A Literature Review on Document Clustering." *Information Technology Journal* 9 (5): 993–1002.

Rego, Lopo L., Matthew T. Billett, and Neil A. Morgan. 2009. "Consumer-Based Brand Equity and Firm Risk." *Journal of Marketing* 73 (6): 47–60.

Richards, Stan. 1998. “Building a Brand.” In *Presentation to the Texas A and M University Center for Retailing Studies Symposium, Dallas, TX*.

Salehan, Mohammad, and Dan J Kim. 2015. “Predicting the Performance of Online Consumer Reviews: A Sentiment Mining Approach to Big Data Analytics.” *Decision Support Systems*.

Salton, Gerard, Anita Wong, and Chung-Shu Yang. 1975. “A Vector Space Model for Automatic Indexing.” *Communications of the ACM* 18 (11): 613–20.

Scheunders, Paul. 1997. “A Genetic c-Means Clustering Algorithm Applied to Color Image Quantization.” *Pattern Recognition* 30 (6): 859–66.

Selim, Shokri Z, and Mohamed A Ismail. 1984. “K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality.” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 1: 81–87.

Shahnaz, Farial, Michael W. Berry, V.Paul Pauca, and Robert J. Plemmons. 2006. “Document Clustering Using Nonnegative Matrix Factorization.” *Information Processing & Management* 42 (2): 373–86.

Silva, Nádia F.F. da, Eduardo R. Hruschka, and Estevam R. Hruschka. 2014. “Tweet Sentiment Analysis with Classifier Ensembles.” *Decision Support Systems* 66 (October): 170–79.

Simon, Carol J., and Mary W. Sullivan. 1993. “The Measurement and Determinants of Brand Equity: A Financial Approach.” *Marketing Science* 12 (1): 28–52.

Song, Wei, Cheng Hua Li, and Soon Cheol Park. 2009. “Genetic Algorithm for Text Clustering Using Ontology and Evaluating the Validity of Various Semantic Similarity Measures.” *Expert Systems with Applications* 36 (5): 9095–9104.

Song, Wei, and Soon Cheol Park. 2006. “Genetic Algorithm-Based Text Clustering Technique: Automatic Evolution of Clusters with High Efficiency.” In *Web-Age Information Management Workshops, 2006. WAIM’06. Seventh International Conference on*, 17–17. IEEE.

Song Wei, and Soon Cheol Park. 2009. “Genetic Algorithm for Text Clustering Based on Latent Semantic Indexing.” *Computers & Mathematics with Applications* 57 (11-12): 1901–7.

Sotiropoulos, D. N., Chris D. Kounavis, Panos Kourouthanassis, and George M. Giaglis. 2014. “What Drives Social Sentiment? An Entropic Measure-Based Clustering Approach towards Identifying Factors That Influence Social Sentiment Polarity.” In *Information, Intelligence, Systems and Applications, IISA 2014, The 5th International Conference on*, 361–73. IEEE.

"Super Bowl Analysis Takes Us Beyond the Tweets." 2014. *A Smarter Planet Blog*. Accessed March 31. http://asmarterplanet.com/blog/2012/02/super-bowl-analysis-takes-us-beyond-the-tweets.html.

Surowiecki, James. 2005. *The Wisdom of Crowds*. Random House LLC.

Swait, Joffre, Tulin Erdem, Jordan Louviere, and Chris Dubelaar. 1993. "The Equalization Price: A Measure of Consumer-Perceived Brand Equity." *International Journal of Research in Marketing* 10 (1): 23–45.

Tagarelli, Andrea, and George Karypis. 2013. "A Segment-Based Approach to Clustering Multi-Topic Documents." *Knowledge and Information Systems* 34 (3): 563–95.

Tirunillai, Seshadri, and Gerard J. Tellis. 2012. "Does Chatter Really Matter? Dynamics of User-Generated Content and Stock Performance." *Marketing Science* 31 (2): 198–215.

Tong, Xiao, and Jana M. Hawley. 2009. "Measuring Customer-Based Brand Equity: Empirical Evidence from the Sportswear Market in China." *Journal of Product & Brand Management* 18 (4): 262–71.

Tou, Julius T, and Rafael C Gonzalez. 1974. "Pattern Recognition Principles." *Pattern Recognition in Physics* 1.

Tsai, Chih-Fong, William Eberle, and Chi-Yuan Chu. 2013. "Genetic Algorithms in Feature and Instance Selection." *Knowledge-Based Systems* 39 (February): 240–47.

Tufekci, Zeynep. 2013. "Big Data: Pitfalls, Methods and Concepts for an Emergent Field." *Methods and Concepts for an Emergent Field (March 7, 2013)*. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2229952.

Tumasjan, Andranik, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." *ICWSM* 10: 178–85.

Vapnik, Vladimir. 2013. *The Nature of Statistical Learning Theory*. Springer Science & Business Media.

Veloutsou, Cleopatra, George Christodoulides, and Leslie de Chernatony. 2013. "A Taxonomy of Measures for Consumer-Based Brand Equity: Drawing on the Views of Managers in Europe." *Journal of Product & Brand Management* 22 (3): 238–48.

Wallace, Byron C. 2015. "Sparse, Contextually Informed Models for Irony Detection: Exploiting User Communities, Entities and Sentiment." In ACL.

Washburn, Judith H., and Richard E. Plank. 2002. "Measuring Brand Equity: An Evaluation of a Consumer-Based Brand Equity Scale." *Journal of Marketing Theory and Practice* 10 (1).

Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. 2005. "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis." In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 347–54. Association for Computational Linguistics.

Wu, Yonghui, Yuxin Ding, Xiaolong Wang, and Jun Xu. 2010. "A Comparative Study of Topic Models for Topic Clustering of Chinese Web News." In *Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on*, 5:236–40. IEEE.

Xu, R., and D. WunschII. 2005. "Survey of Clustering Algorithms." *IEEE Transactions on Neural Networks* 16 (3): 645–78.

Xu, Wei, Xin Liu, and Yihong Gong. 2003. "Document Clustering Based on Non-Negative Matrix Factorization." In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, 267–73. ACM.

Yoo, Boonghee, and Naveen Donthu. 2001. "Developing and Validating a Multidimensional Consumer-Based Brand Equity Scale." *Journal of Business Research* 52 (1): 1–14.

Yu, Yang, Wenjing Duan, and Qing Cao. 2013. "The Impact of Social and Conventional Media on Firm Equity Value: A Sentiment Analysis Approach." *Decision Support Systems* 55 (4): 919–26.

Zeng, Daniel, Hsinchun Chen, Robert Lusch, and Shu-Hsing Li. 2010. "Social Media Analytics and Intelligence." *Intelligent Systems, IEEE* 25 (6): 13–16.

Zhong, Shi, and Joydeep Ghosh. 2005. "Generative Model-Based Document Clustering: A Comparative Study." *Knowledge and Information Systems* 8 (3): 374–84.

Zwass, Vladimir. 2010. "Co-Creation: Toward a Taxonomy and an Integrated Research Perspective." *International Journal of Electronic Commerce* 15 (1): 11–48.

Appendix 1

Sample Tweets for AT&T

	AT&T sucks... Lets tell our customers something is unlimited then charge them outrageous overage fees.
	@DaGreatNatePeck are you serious?! Fml. I hate AT&T so much.
	Fck AT&T n they high ass bill
	Who else thinks the AT&T commercials with the little kids are really annoying?
	when AT&T texts me a survey to do #no
	AT&T wanna charge me $200 for a replacement phone
	If I have to go another day with this no cell service bullshit Im going to AT&T headquarters and kicking the first person I see in the face
	"RT @BobbyBellafonte: I hate AT&T smfh they should stand for (A)ssholes (T)rying (T)otakeyourmotherfuckingmoney"
	AT&T gonna make me punch them in the face.
	No AT&T I dont feel like talking about how my day is.
	My mom yelling at AT&T right now>>>>>
	The AT&T commercials with the kids in the classroom are pretty funny
	These AT&T commercials with the kids are all hilarious!
	Funniest AT&T ad ever
	I love the AT&T commercials lol
	Thank god for AT&T I got all the movie channels
	That new AT&T commercial is classic! Haha make my day
	I love all the at&t commercials with the little kids theyre so funny to me
	I love those AT&T commercials with Kevin Durant.
	Thank you AT&T for allowing me to have service in MiddleOfNoWhere, Louisiana
	RT @FetchKenBarbie: At AT&T getting my iPhone 5!!!!!
	AT&T has the best commercials #nodoubt #handsdown
	Thank you AT&T for your service, that ugly ass number is finally blocked.
	Damn Verizon wanted 400 just to start an account not happening Im with AT&T lol
	AT&T has the best service!! Idc

Sample Tweets for Verizon

	Whats up with Verizons service? Its been horrible here lately. #FixIt
	Verizon needs to get their stuff together today.
	Fuck you Verizon, for real.
	Verizon can suck my daddy dick w/ those data alert messages ! 50% my ass
	Ughh I really hate Verizon !
	YOU SUCK VERIZON.
	I hate Verizon fios I cant find any of my channels
	Agh I need help with my phone and I want to talk to an actual person but thats impossible with Verizon #pissed
	WHAT THE HECK VERIZON
	@OllieRoo *Stuck in a Verizon lab till Thursday*
	I FUCKING HATE VERIZON FIOS!!!!!
	Now my phone isnt working again what is up Verizon ;-;
	Verizon LTE is fucking up.
	WHAT THE HECK VERIZON
	The verizon guy was supposed to be here from 8-10. Where dafuq is he
	I hate Verizon!!
	Glad I got Verizon
	I love Verizon.! Having a signal where ever I go>>>>>>> #SprintWho
	That guy at the Verizon store was a cutie >>>>
	Verizon The Best Cable lol
	RT @MsNewNew4u: I love Verizon
	@lightskin_sweet and I want Verizon, but its a rap now lol
	I need a Verizon iPhone !
	Verizon 3G. Praise the lerd
	@VerizonSupport wow you guys are the best. But I forgot I had Comcast here and not Verizon. Thanks anyway
	we just switched to Verizon so now we have the channel that shows the Brits
	Verizon might b high but its well worth it
	I freaking LOVE Verizon!!!!! This lady just handled my dirty work like she was my homegirl

AT& T Sentiment Classification Results

10-fold cross-validation results									
Accuracy per fold									
0,7431 694	0,7704 918	0,7814 2077	0,7978 1421	0,8032 7869	0,765 02732	0,8131 8681	0,847 45763	0,7912 0879	0,7692 3077
Mean Accuracy: 0,788 (+/- 0,053)									
Precision per fold									
0,8169 0141	0,8	0,8354 4304	0,858 97436	0,8701 2987	0,847 22222	0,862 5	0,8795 1807	0,8461 5385	0,7976 1905
Mean Precision: 0,841 (+/- 0,054)									
Recall per fold									
0,630 43478	0,73118 28	0,709 67742	0,7204 3011	0,7204 3011	0,655 91398	0,75	0,7934 7826	0,7173 913	0,7282 6087
Mean Recall: 0,716 (+/- 0,086)									
F-score per fold									
0,7116 5644	0,7640 4494	0,7674 4186	0,7836 2573	0,7882 3529	0,7393 9394	0,802 32558	0,8342 8571	0,7764 7059	0,7613 6364
Mean F-score: 0,773 (+/- 0,064)									

Verizon Sentiment Classification Results

10-fold cross-validation results									
Accuracy per fold									
0,728 87324	0,71126 761	0,7218 3099	0,7112 6761	0,7067 1378	0,7491 1661	0,7243 8163	0,71731 449	0,7243 8163	0,7349 8233
Mean Accuracy: 0,723 (+/- 0,024)									
Precision per fold									
0,706 52174	0,6571 4286	0,695 65217	0,6542 0561	0,656 86275	0,71153 846	0,7176 4706	0,6635 5140	0,698 92473	0,7127 6596
Mean Precision: 0,687 (+/- 0,050)									
Recall per fold									
0,565 21739	0,6000 0000	0,556 52174	0,608 69565	0,5826 0870	0,6434 7826	0,5304 3478	0,6173 9130	0,5652 1739	0,5826 0870
Mean Recall: 0,585 (+/- 0,063)									
F-score per fold									
0,628 01932	0,6272 7273	0,6183 5749	0,6306 3063	0,6175 1152	0,6757 9909	0,6100 0000	0,639 63964	0,6250 0000	0,6411 4833
Mean F-score: 0,631 (+/- 0,035)									

Appendix 2

Indicative Tweets for UBER per topic and per cluster

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 center 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*

Sentiment Classification Results

10-fold cross-validation results									
Accuracy per fold									
0,832 40223	0,8379 8883	0,9385 4749	0,876 40449	0,966 29213	0,9775 2809	0,7977 5281	0,864 40678	0,8474 5763	0,920 90395
Mean Accuracy: 0,886 (+/- 0,116)									
Precision per fold									
0,858 82353	0,907 89474	0,944 44444	0,9367 0886	1,00	0,978 02198	0,9230 7692	0,875	0,8387 0968	0,8725 4902
Mean Precision: 0,914 (+/- 0,100)									
Recall per fold									
0,8021 978	0,7582 4176	0,9340 6593	0,8131 8681	0,9340 6593	0,978 02198	0,6593 4066	0,8555 5556	0,866 66667	0,988 88889
Mean Recall: 0,859 (+/- 0,198)									
F-score per fold									
0,829 54545	0,8263 4731	0,9392 2652	0,8705 8824	0,965 90909	0,978 02198	0,769 23077	0,8651 6854	0,8524 5902	0,9270 8333
Mean F-score: 0,882 (+/- 0,129)									