Machine Learning Applications in Social Media Analytics: A State-of-Art Analysis

Sosyal Medya Analitiğinde Makine Öğrenmesi Uygulamaları: Literatür İncelemesi

Birce DOBRUCALI, İzmir University of Economics, Turkey, birce.dobrucali@ieu.edu.tr

Orcid No: 0000-0003-3462-0606

Burcu ILTER, Dokuz Eylül University, Turkey, burcu.ilter@deu.edu.tr

Orcid No: 0000-0002-3781-7263

Abstract: Social media analytics (SMA), referring to the collection and analysis of user generated data from social media platforms, attract attention of both researchers and practitioners striving to derive consumer insights. The SMA domain grows multifariously, with a highlight on the capability of machine learning algorithms in capturing noteworthy insights through processing high-volume and complex data in a cost effective way. As machine learning applications draw attention as a fertile area that may re-shape the future of SMA, there is a need to comprehend trends and approaches in an integrative framework. Accordingly, this study aims to present an integrative framework by portraying machine learning application trends and approaches in SMA. 42 scientific articles published in refereed scientific business, management, and computational science journals between the years 2013 and 2019 are analyzed via systematic literature review based on visual text mining method (SLR-VTM). The results revealed five distinctive research clusters as: (1) review sites, (2) microblogs, (3) social networking sites, (4) content communities, (5) cross-media. This analysis plays a crucial role for enhancing our understanding regarding the intellectual structure of the field, acknowledging the leading studies of the domain, better positioning future research, and determining gaps and new paths for researchers.

Keywords: Social Media Analytics, Machine Learning, User-Generated Content

JEL Classification: M10, M30, M31

Öz: Sosyal medya platformlarından kullanıcı tarafından oluşturulan verilerin toplanması ve analiz edilmesini ifade eden sosyal medya analitiği (SMA), tüketici içgörüleri elde etmeye çalışan araştırmacıların ve uygulayıcıların ilgi odağındadır. Bu alan, makine öğrenimi algoritmalarının yüksek hacimli ve karmaşık verileri uygun maliyetli bir şekilde işleyerek kayda değer içgörüleri yakalama kapasitesine paralel olarak çok yönlü bir şekilde büyümesini sürdürmektedir. Makine öğrenimi uygulamaları, sosyal medya analitiğinin geleceğini yeniden şekillendirebilecek verimli bir alan olarak dikkat çektiğinden, mevcut trendleri ve yaklaşımları bütünleştirici bir çerçevede anlamaya ihtiyaç vardır. Bu bağlamda, mevcut çalışma sosyal medya analitiği alanındaki makine öğrenimi uygulama trendlerini ve yaklaşımlarını bütünleştirici bir çerçevede sunmayı amaçlamaktadır. 2013-2019 yılları arasında hakemli bilimsel dergilerde yer alan ve işletme, yönetim ve bilgisayar bilimleri alanında yayınlanan 42 bilimsel makale, görsel metin madenciliğine dayalı sistematik literatür taraması yöntemi ile analiz edilmiştir. Sonuçlar beş farklı araştırma kümesini ortaya çıkarmıştır: (1) inceleme siteleri, (2) mikrobloglar, (3) sosyal ağ siteleri, (4) içerik toplulukları, (5) platformlar arası çalışmalar. Mevcut çalışma, alanın entelektüel yapısı hakkındaki anlayışı geliştirmek, alanın önde gelen çalışmalarına dikkat çekmek, gelecekteki araştırmaların daha iyi konumlandırılmasına yönelik olarak alandaki boşlukları ve yeni araştırma alanlarını belirlemek açısından önemli bir rol oynamaktadır.

Anahtar Kelimeler: Sosyal Medya Analitiği, Makine Öğrenmesi, Kullanıcı Katkılı İçerik

JEL Sınıflandırması: M10, M30, M31

Makale Geçmişi / Article History

Başvuru Tarihi / Date of Application	: 10 Şubat / February 2020
Kabul Tarihi / Acceptance Date	: 8 Aralık / December 2020

1. Introduction

Social media networks have become a crucial part of consumers' daily lives (Shareef et al., 2018) and re-shaped the way of information acquisition, opinion sharing, and communication (Kapoor et al., 2018). These platforms, driven by user-generated content (i.e. Twitter, Facebook, Instagram, Snapchat, LinkedIn, Whatsapp), are highly influential for a wide range of settings including consumer and seller behavior, entrepreneurship, political issues, and venture capitalism (Greenwood and Gopal, 2015). The number of social media users upsurged to 3.484 billion by 2019, and the average number of social media accounts per person reached 8.9 with the average use of 2 hours 16 minutes per day (Chaffey, 2019). In parallel with the extensive daily usage and ascending number of active users, social media communication grabbed attention of both researchers and practitioners due to its ability of enhancing product/service acceptability and creating group opinion (Shareef et al., 2018). Moreover, both commercial and non-profit organizations are motivated to comprehend issues and trends in social media as the proxy for identifying related risks and changes in the communication, and deriving insights (Stieglitz et al., 2018). The extensive usage of social media has led to accumulation of large sets of data in various formats (i.e. text, pictures, videos, sounds, geotags), which has been termed as Social Media Big Data (Stieglitz et al., 2018) and characterized as being high in volume, velocity, and variety which make the large set of information to require cost-effective and innovative methods of data processing (Beyer and Laney, 2012). The user generated content on social media platforms produces six main types of social media big data: (1) Service data - Data provided to social media platform for signing up (i.e. name, e-mail, age etc.), (2) Disclosed data - Content posted on users' own pages (i.e. messages, comments, photos), (3) Entrusted data – Content posted on another user's page, (4) Incidental data – Content other users post about another user, (5) Behavioral data – Data the social media platform collects for tracking your usage habits, (6) Derived data – Data about a user that is derived from all other types of data (Schneider, 2010). The collected social media big data is then being utilized by various parties (i.e. governments, companies, social media platforms, marketing agencies) with diverse aims. The main utilization fields of social media networking data can be listed as: determining sales and marketing strategies, identifying issues discussed by countries, determining and executing government policies, analyzing user profiles and personalities, developing recommendation systems, resolving and clarifying crimes, evaluating educational parameters, analyzing emotion and sentiment, producing personalized advertisements, extracting incomplete information, and analyzing social relationships (Demirci and Sağıroğlu, 2017). Variety, volume, and criticality of social media big data for various fields yield both a powerful and cost-effective analysis technique. Accordingly, with its ability to process high-volume and complex data, machine learning applications has started to become prevalent in social media analytics, and draw attention as a fertile area that may re-shape the future of research on social media analytics. The noteworthy contribution of machine learning applications to research in the field of social media analytics makes the domain more alluring, since these applications are highly effective, far-reaching, and fruitful for deriving insights from social media big data.

Due to multifaceted and fertile nature of the social media analytics, researchers from various domains, including different branches of social sciences and computational sciences, directed attention to the field. Regardless of progressive research on social media analytics through machine learning applications in different research domains, efforts for accumulating and mapping the extant body of knowledge have been restricted. Accordingly, the aim of this study is to present an integrative framework by portraying machine learning application trends and approaches in SMA to enhance our understanding regarding the intellectual structure of the field, and acknowledge the leading studies of the domain. To best of our knowledge, there exist a lack in literature regarding a holistic framework of machine learning applications in social media analytics. However, previously mentioned extensive and critical utilization scope of the domain urges for an integrative framework. This paper is an attempt to review a broad range of literature on the employment of various machine learning algorithms for deriving valuable information from social media big data. Using systematic literature review based on visual text mining method (SLR-VTM), our paper focuses on the following research question: What is the current intellectual structure of machine learning applications in social media analytics research and how has it evolved? This includes identification of the main topics studied and the main clusters of research present in the field.

The remainder of the paper is organized as follows. First, the followed methodological procedures, and the results of the co-occurrence analysis are presented. In the next section, detailed analysis of studies under each cluster is presented. Later, evaluations and discussion on the findings of the review are provided. Finally, the article is concluded by providing a summary of this research and highlighting some avenues for future research.

2. Research Methodology

In interdisciplinary research domains, traditional literature reviews remain insufficient in terms of review scope and are prone to judgement bias (Gurzki and Woisetschläger, 2017), whereas text-mining approaches offer a way for overcoming such limitations (Porter et al., 2002). Combination of these complementary methods provide a robust analysis of research streams and key concepts (Randhawa et al., 2016). Accordingly, Systematic Literature Review based on Visual Text Mining (SLR-VTM) methodology enhances performance and efficiency of such efforts (Felizardo et al., 2011; Fabbri et al., 2012; Mergel et al., 2015). Hence, via employing VOSViewer as a visual text mining tool, this study employs SLR-VTM approach for presenting multilevel linkages between different domains and mapping various research steams.

As of the first step, research streams in the domain are determined by clustering the extant body of knowledge in the field through VOSViewer, which is used to visualize similarities and patterns in a given data set (Bragge et al., 2019). Thereafter, since systematic review enables integration of various studies on a subject through summarizing common elements while contrasting the differences (Meredith, 1993), the structure and characteristics of each cluster is examined by reviewing the studies systematically.

2.1. Data

The raw data consisting of publications is obtained through the core collection of Thomson Reuter’s ISI Web of Knowledge database, which comprises an extensive range of scholarly publications, and enables access to authorship, affiliation and citation information of publications. Moreover, even though various alternatives exist (i.e. Scopus, Google Scholar etc.), Thomson Reuter’s ISI Web of Knowledge is being used by the majority of the scholars for benchmarking analyses, bibliometric researches and reviews (Harzing, 2013). Hence, for ensuring the scientific quality and compatibility of the obtained information with VOSviewer, ISI Web of Knowledge is selected as the data source.

For article filtering, two set of search terms were determined in compliance with the context of the review. The first group of search terms, which aim ensuring the congruency of data analysis techniques with the purpose of this study, consists of following keywords: “machine learning”, “topic modelling”, “sentiment analysis”, “natural language processing”, “text analytics”, “neural networks” and “artificial neural networks”. The second group of search terms aim assuring that employed data is gathered from social media platforms, and consists of the subsequent terms: “social media”, “social media analytics”, “user generated content”. Following this step, for a further refinement, publication selection procedure was conducted based on subsequent eligibility criteria: (1) to investigate a phenomena related with social media analytics; (2) to hold an empirical nature and entail social media data; (3) to employ machine learning applications for data analysis; (4) to be published (articles in “forthcoming” status is not included) in international journals, and not in books, conference proceedings, or research reports.

Table 1. Employed Keywords

Keywords for ensuring the congruency of data analysis techniques with the purpose of the review	**Keywords for assuring that employed data is gathered from social media platforms**
Machine Learning Topic Modelling Sentiment Analysis Natural Language Processing Text Analytics Neural Networks Artificial Neural Networks	Social Media Social Media Analytics User-Generated-Content

At the end of the practical screening process through application of afore-mentioned search terms and research criteria, a total of 42 publications were obtained for further analysis. The earliest date of selected publications is 2013, whilst the latest ones are 2019. The majority of the studies were published in business and marketing related journals, while the remaining were in computational science related journals.

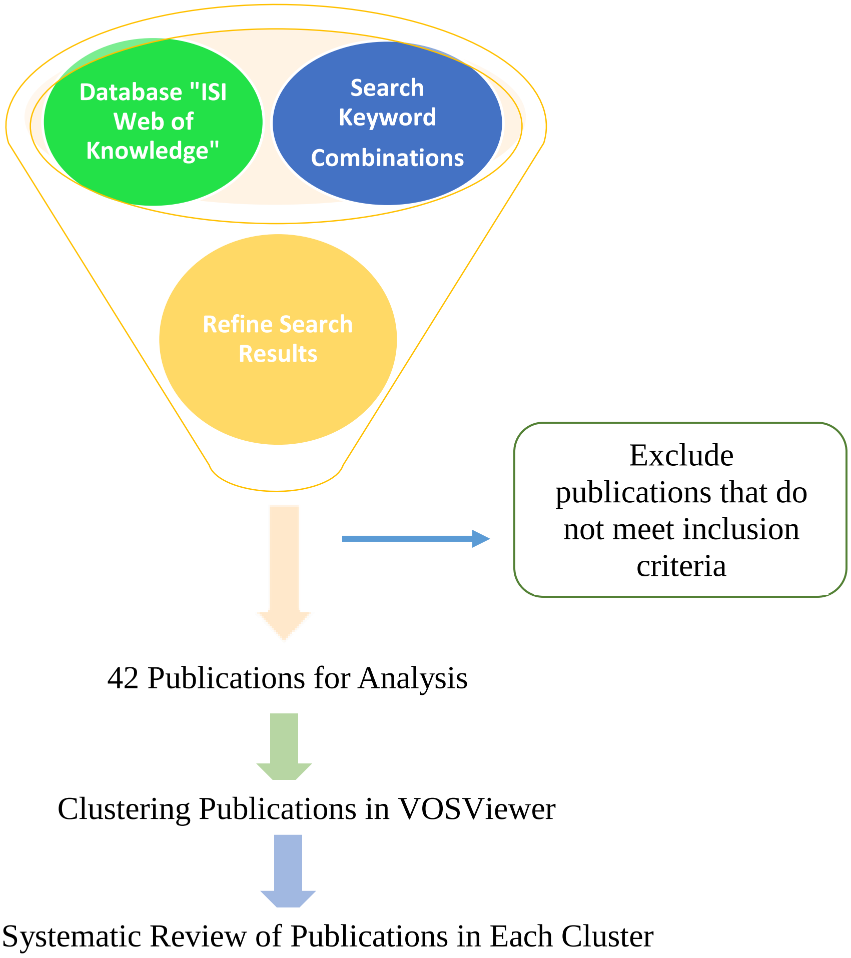
Database "ISI Web of Knowledge"
Search Keyword Combinations
Refine Search Results
Exclude publications that do not meet inclusion criteria
42 Publications for Analysis
Clustering Publications in VOSViewer
Systematic Review of Publications in Each Cluster

Figure 1. Roadmap for the Selection of Publications and Review Process

2.2. Descriptive Review of the Literature

Articles that integrate machine learning applications to social media analytics domain demonstrate a growing trend as indicated in Fig.1, which illustrates publication frequency available in the WoS database. Among the filtered and refined publications available in the WoS database, a stagnancy phase is spotted between the years 2013 and 2016. From this point onwards, the number of publications in the related field increased considerably with a peak in years 2017 and 2018. Since ISI Web of Knowledge is being taken as a research quality benchmark for the majority of the scholars (Harzing, 2013), by depending on the increasing number of published articles in that database, it may be stated that qualified research efforts in the domain has accelerated. Accordingly, such an acceleration is expected to bring new dimensions to the stream.

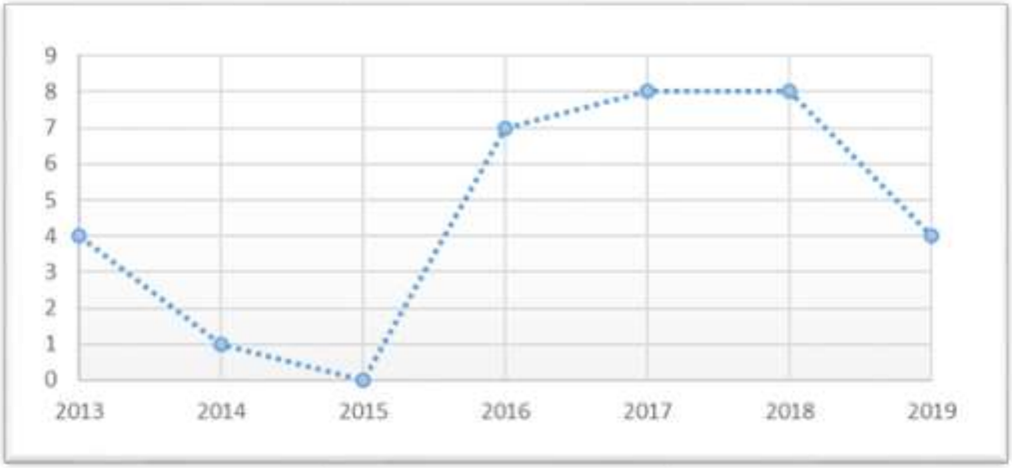

Figure 2. Article Frequency per Year

The majority of the articles (around 52%) were mainly published in business, marketing, and management oriented publication outlets. As indicated in Table 1, within the specified domain, marketing journals outweigh business and management journals. Among business, marketing, and management oriented journals, Journal of Retailing and Consumer Services, Journal of Hospitality and Marketing Management, Psychology and Marketing, and Tourism Management provided the highest number of articles included in this review. From computational science oriented publication outlets, Expert Systems with Applications, Applied Soft Computing, and Knowledge-Based Systems are the journals with the highest contribution rate.

Table 2. Contributing Journals

Business, Marketing and Management Oriented Publication Outlets	**No. of Articles**	**Weight 52.36%**
Journal of Retailing and Consumer Services	3	7.14%
Journal of Hospitality Marketing & Management	3	7.14%
Psychology and Marketing	3	7.14%
Tourism Management	3	7.14%
International Journal of Information Management	1	2.38%
Journal of Travel and Tourism Marketing	1	2.38%
European Journal of Marketing	1	2.38%
Journal of Interactive Marketing	1	2.38%
Journal of Consumer Marketing	1	2.38%
Management Science	1	2.38%
Journal of Business Research	1	2.38%
International Journal of Research in Marketing	1	2.38%
Marketing Science	1	2.38%

Journal of Political Marketing	1	2.38%
Computational Science Oriented Publication Outlets		**47.64%**
Expert System with Applications	3	7.14%
Applied Soft Computing	3	7.14%
Knowledge-Based Systems	3	7.14%
Decision Support Systems	2	4.76%
The Electronic Library	1	2.38%
J of Computational Science	1	2.38%
Information Systems Research	1	2.38%
Program	1	2.38%
BMJ Quality & Safety	1	2.38%
Online Information Review	1	2.38%
Omega	1	2.38%
Focus on Information Technology	1	2.38%
Neural Networks	1	2.38%

The vast majority of the reviewed articles (56%) conducted analysis by employing qualitative data, and the greater part of those are published in computational science oriented outlets as indicated in Table 2. A relatively smaller set of studies (36%) conducted analysis with both qualitative and quantitative data, and the majority of those are published in business, management and marketing oriented outlets. Only three studies conducted analysis uniquely with quantitative data, and one study conducted analysis with visual data.

Table 3. Type of Employed Data

Business, Marketing and Management Oriented	**Computational Science Oriented**
Qualitative	
Rogers et al. (2017)	Ghiassi et al. (2013)
Dhaoui et al. (2017)	Ikeda et al. (2013)
Calheiros et al. (2017)	Jin-Jang et al. (2013)
Xiang et al. (2017)	Schniederjans et al. (2013)
Nave et al. (2018)	Vazquez et al. (2014)
Bilro et al. (2018)	Ali et al. (2016)
Park et al. (2018)	Tripathy et al. (2016)
İlhan et al. (2018)	Hawkins et al. (2016)
Klostermann et al. (2018)	Kim et al. (2016)
	Pournarakis et al. (2017)
	Wong et al. (2017)
	D'Avanzo (2017)
	Rathan et al. (2018)
	Klostermann et al. (2018)
	Pelaez et al. (2019)

Quantitative	
Liu et al. (2016) Bigne et al. (2019)	Nilashi et al. (2018)
Qualitative & Quantitative	
Agnihotri & Bhattacharya (2016) Moro et al. (2016) Walker et al. (2017) Ryoo & Bendle (2017) Pineiro-Chousa et al. (2017) Heng et al. (2018) Eslami & Ghasemaghaei (2018) Lee et al. (2018) Costa et al. (2019)	Wang et al. (2019) Kwok & Yu (2013) Hu et al. (2018) Adamopoulos et al. (2018) Nguyen et al. (2018) Jimenez-Marquez et al. (2019)
Visual	
Giglio et al. (2019)	

3. Evaluations and Discussion

3.1. Research Clusters

For depicting the intellectual structure of the domain and identifying research clusters, co-occurrence analysis of VOSViewer was employed. The analysis was concluded in a map indicating numbers of co-occurrences of the terms appearing in the sample publications' titles and abstracts. The result of the co-occurrence analysis is shown in Fig. 2.

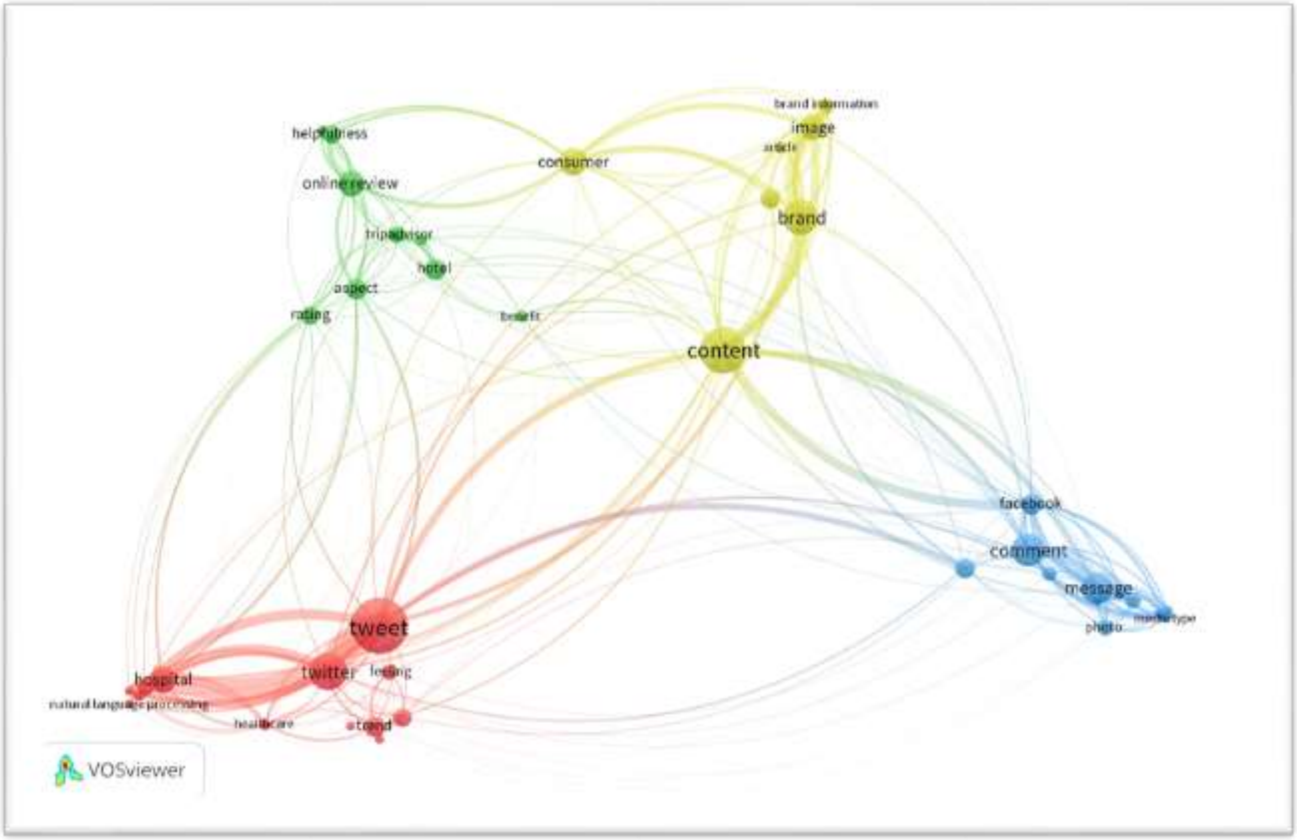

Figure 3. Term Co-occurrence Map

The terms shown in the Fig. 2 appear at least four times in diverse publications partaking in the sample. Yet, it should be noted that not all terms are shown as labels for evading clutter. The size of bubbles and labels represents frequency, as the terms with larger bubbles and fonts appear more frequently in the sample. Additionally, the proximity of terms and links depict frequency of the terms within same articles.

Explicating the term co-occurrence map, the current literature consists of four main interconnected research clusters:

1. Review Sites *(green cluster)* – i.e. Yelp, Tripadvisor, IMDb etc.
2. Microblogs *(red cluster)* – i.e. Twitter, Tumblr, Foursquare etc.
3. Social Networking Sites *(blue cluster)* – i.e. Facebook, Whatsapp, WeChat etc.
4. Content Communities *(yellow cluster)* – i.e. Instagram, Youtube, Flickr etc.

In addition to clusters revealed by term co-occurrence map, as an output of the further review analysis, a fifth group of research is added in findings section for covering studies that employ cross-media data. Studies reviewed under this heading conduct social media analytics through employing data obtained from multiple types of social media networks simultaneously (i.e. both Facebook and Twitter data), and for preventing a possible conceptual ambiguity this steam of research is analyzed separately from the remaining research.

3.1.1. Cluster 1 - Review Sites

The first cluster contains “online review”, “Tripadvisor”, “hotel”, “rating”, “aspect”, “benefit” and “helpfulness” as the main terms. Studies belonging this cluster conducted social media analytics through analyzing data obtained from review sites including Tripadvisor, Amazon reviews, Yelp, and IMDb. Among fourteen studies, eight are published in business and marketing related journals, while the remaining are in computational science related journals. The earliest date of publications belonging to this cluster is in 2016, whilst the latest ones are in 2019. Detailed information about the publications included in this cluster is provided in Table 4.

Table 4. Detailed Information Regarding the Publications in Review Sites Cluster

Nave et al. (2018)	Employing text mining and sentiment analysis techniques for structuring online reviews to be used in a decision support system.	Correlated Topic Models	Most reviews have a positive tone. Words like food and place are found to be the most positive ones, indicating that consumers consider location as a vital driver of satisfaction.
Bilro et al. (2018)	Analyzing dimensions of online customer engagement, as well as their involvement, emotional states, experience and brand advocacy in online reviews.	Sentiment: WordNet2.1, Engagement Dictionary	Results indicate high influence of the engagement cognitive processing and hedonic experience on customers’ review effort. Customers are more engaged in positively advocating a company/brand than the contrary.
Agnihotri & Bhattacharya (2016)	Exploring the impact of review readability and sentiment on the helpfulness of the review.	LIWC	Reviewer experience heuristically influences consumers’ trust of online reviews, thus making even too simplistic or extremely sentimental reviews helpful.

Heng et al. (2018)	Revealing the factors that impact consumers’ purchase decisions of different brands of coffee.	N/A	Amazon service, physical feature, flavor feature, and subjective expression impact the helpfulness of customer reviews. Readers find objective reviews more helpful, and helpfulness has a concave relationship with the text length.
Eslami & Ghasemaghaei (2018)	Investigating (1) review positiveness' impact on product sales, (2) online review score inconsistency's impact on sales, (3) the extent to which the impacts of online review positiveness and review score inconsistency on product sales are different between high and low involvement products.	Sentiment: R in the Comprehensive R Archive Network1 (CRAN)	For high involvement products, review text sentiment, review score and review score inconsistency impact product sales. For low involvement products review title sentiment, and review score impact product sales.
Costa et al. (2019)	Predicting whether a review was incentivized based on various review features.	Sentiment: VADER, Random Forest	The generated model is able to correctly predict bias in a review over 75% of times, based on some characteristics like the length of a review, the helpfulness rate and the sentiment polarity scores. The most important variable, is the number of characters used. The most critical sentiment-related variable is the overall compound score, which was higher on the incentivized reviews.
Rathan et al. (2018)	Proposing a feature level sentiment analysis framework for Twitter including features of emoji detection, spelling correction and emoticon detection.	SVM	The proposed model provided good accuracy in sentiment detection.

Pelaez et al. (2019)	Developing a decision-making model that creates alternative ranking by contextualizing unsolicited opinions expressed by decision-makers in social media.	N/A	The proposed model yielded an alternative ranking that contextualizes the feelings/opinions represented through feeling consistent with data extracted from social media.
Nilashi et al. (2018)	Developing a hotel recommendation framework via employing multi-criteria ratings.	SOM, EM, CART	Developed method successfully employs online reviews for making precise recommendations in TripAdvisor.
Hu et al. (2018)	Predicting box-office performance via combining movie information, external and review factors, and determining the most crucial factor of influencing performance.	Three DM techniques employed: M5 model tree, support vector regression and linear regression	The number of reviews has the greatest impact on box-office performance, and review content impacts box-office performance only for specific movie genres.
Tripathy et al. (2016)	Classifying movie reviews using various supervised ML algorithms.	NB, Maximum Entropy, Stochastic Gradient Descent, SVM	As the n value in n-gram increases, accuracy decreases. Use of unigram, bigram, trigram and combinations yield better results.
Jimenez-Marquez et al. (2019)	to present a computational framework for the management of BD focusing primarily, but not exclusively, on sets of information containing UGC.	NLTK and Pandas packages; Multi-layer Perceptron, Support Vector Classifier, Logistic Regression, Linear Support Vector Classification, Linear classifier with Stochastic Gradient Descent training, and Naive Bayes.	Various algorithms were trained for two, three and five classes, with the best results being found for binary classification.
Calheiros et al. (2017)	Classifying online hotel reviews based on their sentiment.	LDA	Hotel food generates ordinary positive sentiments, while hospitality generates both ordinary and strong positive feelings.

Ali et al. (2016)	Proposing a classification technique for feature review's identification and semantic knowledge for opinion mining.	SVM (for eliminating irrelevant reviews) & FDO (for computing polarity)	Proposed classification technique increases accuracy of review polarity calculation from %70s to %80s

Note: “N/A” abbreviation, which stands for not applicable, is used for the publications that did not indicate the employed algorithm and/or lexicon.

Five studies belonging to this cluster (Agnihotri and Bhattacharya, 2016; Eslami and Ghasemaghaei, 2018; Heng et al., 2018; Rathan et al., 2018; Costa et al., 2019) employed Amazon reviews mainly for conducting sentiment analysis. Uniquely Heng et al. (2018) applied topic modelling approach by employing linear discriminant analysis (LDA) for determining helpfulness of reviews on consumers’ coffee purchase decisions, and identified four factors (Amazon service, physical feature, flavor feature, and subjective expression) that impact the helpfulness of customer reviews. Remaining four studies conducted sentiment analysis through employing various statistical learning functions and packages for discovering diverse relationships. Similar to Heng et al. (2018), Agnihotri and Bhattacharya (2016) conducted a research on review helpfulness and employed over 2.000 Amazon reviews about technological products for exploring impact of readability and sentiment polarity on helpfulness of the review. Sentiment analysis was conducted by LIWC (Linguistic Inquiry and Word Count) package, and it was found that being moderated by the reviewer experience, review content and implied sentiment are key predictors of review helpfulness. Costa et al. (2019) employed over 100.000 Amazon reviews and conducted sentiment analysis via VADER (Valence Aware Dictionary and Sentiment Reasoner) library for anticipating if a review on published book/electronics was incentivized, and generated decision tree models were capable of correctly predicting bias reviews over 75% of times, based on some characteristics including length of review, helpfulness rate and the sentiment polarity. Eslami and Ghasemaghaei (2018) aimed to investigate impact of review positiveness and online review score inconsistency on product sales by analyzing over 15.000 Amazon reviews on musical instruments and digital music through Comprehensive R Archive Network (CRAN), and concluded that for high involvement products, review sentiment, review score and score inconsistency impact sales, and for low involvement products review sentiment, and review score impact sales. Finally, Rathan et al. (2018) developed a sentiment analysis model through Support Vector Machine (SVM) by analyzing

400.000 Amazon reviews on phones and developed model provided 80% accuracy in detecting sentiment.

Three studies (Bilro et al., 2018; Nave et al., 2018; Jimenez-Marquez et al., 2019) used reviews on Yelp for conducting sentiment analysis. Nave et al. (2018) employed 12.000 Yelp reviews on various service providers, conducted sentiment analysis through Semantria software, and analyzed them by employing Correlated Topic Models (CTM) for creating a data-driven decision support system which enables managers to analyze reviews from different perspectives. Bilro et al. (2018) analyzed dimensions of online consumer engagement in online reviews through analyzing 15.000 Yelp reviews on various service providers by employing WordNet2.1 software, and found that engagement cognitive processing dimension and hedonic experience have a high impact on consumers' review endeavor. Lastly, Jimenez-Marquez et al. (2019) aimed presenting a framework for big data management by focusing primarily on sets of information containing Yelp reviews about hotels through several machine learning algorithms, and found that binary classification provides the best results with 88% accuracy.

Hu et al. (2018) and Tripathy et al. (2018) both conducted sentiment analysis through employing movie reviews obtained from IMDb. Hu et al. (2018) investigated whether reviews impact the desire of readers to watch movies, and reviews to predict the movie box-office performance, and concluded that movie reviews enhance the accuracy of box-office performance predictions while highlighting that the performance would remain satisfactory as the quantity of movie reviews remained high, even if the content was negative. Tripathy (2018) aimed classifying review sentiments via several supervised ML algorithms; Naïve Bayes, Maximum Entropy, Stochastic Gradient Descent, Support Vector Machine, and concluded that as the n value in n-gram increases, accuracy declines.

Calheiros et al. (2017) and Nilahsi et al. (2018) conducted their research through employing Tripadvisor reviews. Calheiros et al. (2017) employed online hotel reviews in order to classify them by using topic modelling through Linear Discriminant Analysis (LDA), and concluded that hotel food causes ordinary positive sentiments, while hospitality causes both ordinary and strong positive feelings. Nilashi et al. (2018) aimed to use multi-criteria ratings (including overall rating, value, rooms, location, cleanliness, check in/front desk and service) for developing a novel hotel recommendation method, and used various algorithms such as Self-Organizing Map (SOM), Expectation Maximization (EM), Classification and Regression Trees (CART) and fuzzy-based ML techniques for reaching to the most accurate recommendation method. Results of the study confirmed that analysis of online reviews through Fuzzy Rule-Based Ensemble and EM Ensemble leads to precise recommendations in TripAdvisor with 94.5% accuracy.

Even though the first research belonging to this cluster was published in 2016, it gained a considerable acceleration within only few years, indicating that the user comments on the review sites grabbed the most attention from the researchers in the field. The vast majority of the reviewed studies (33%), employed consumer reviews obtained from diverse review sites, and mainly conducted sentiment analysis for gaining in depth understanding on the helpfulness and sentimental tendency of the customer reviews on consumed goods and services from various perspectives. The most examined social media platform of the category is Amazon, and it is followed by Tripadvisor and Yelp. It should also be noted that almost each single study employed a different statistical learning function or software package leading us to the conclusion that a comparison of each function and/or package for diverse research aims remains as a fruitful gap to be investigated. Moreover, findings indicate that consumer reviews on goods compelled more attention when compared with the services. However, when the extensive usage of social media platforms for evaluating consumed services is considered, amplifying the extant knowledge on service evaluation dynamics of consumers may be beneficiary for both researchers and practitioners.

3.1.2. Cluster 2 – Microblogs

The second cluster, which contains studies conducting social media analytics through employing data gathered from microblogs, holds main terms of “tweet”, “Twitter”, “trend” and “natural language processing”. Studies belonging to this cluster conducted social media analytics through analyzing data obtained from microblogs including Twitter and Sina Weibo. Among twelve studies, eight are published in computational science related journals, while the remaining are in business and marketing related journals. The earliest date of publications belonging to this cluster is in 2013, whilst the latest ones are in 2019. Detailed information about the publications included in this cluster is provided in Table 5.

Table 5. Detailed Information Regarding the Publications in Microblogs Cluster

Pournarakis et al. (2017)	Building a genetic algorithm for clustering of tweets in semantically coherent groups.	Sentiment: SVM; LDA topic modelling	The model generates 4 key insights, with regard to (1) overall volume of tweets and prevailing topics discussed, (2) daily sentiment towards the brand, (3) optimal clustering of tweets in semantically-coherent clusters under a distinct topic, and (4) an overall average sentiment assessment of each topic per day.
Wang et al. (2009)	Quantifying the optimal level of marketing aggressiveness in social media for reaching maximum popularity.	Naïve Bayes	The relationship between the level of marketing aggressiveness and popularity follows an inverted U-shaped curve.
D'Avanzo (2017)	Introducing a framework for monitoring Twitter's sentiment regarding Google trends, and enabling users to employ geographic location for monitors.	Four sentiment analysis tools: A Naïve Bayes detection algorithm, a simple voter algorithm, an NLTK-based sentiment analysis algorithm, and a commercial sentiment analysis tool.	Three monitoring tasks (consumer electronics, healthcare, and politics) validated the plausibility of the proposed approach in measuring social media sentiments and emotions regarding the trends emerged on Google searches.

Adamopoulos et al. (2018)	Investigating whether personality traits of online users accentuate or attenuate the effectiveness of WOM in social media.	LDA	Reading WOM messages from similar users in terms of personality increases the probability of a post purchase by 47.58%. On the contrary to extroverts, introverts are responsive to WOM. Agreeable, conscientious, and open social media users are more effective disseminators of WOM. WOM from users with low levels of emotional range affects similar users. A WOM message from an extrovert user to an introvert increases the likelihood of a purchase by 71.28%.
Rogers et al. (2017)	Observing Double Jeopardy and Negative Double Jeopardy in online microblogging environment.	For sentiment: Naive Bayes; For clustering: Fuzzy c-means	Bigger brands suffer from an increased negativity amongst the larger proportion of tweets about them.
Park et al. (2018)	Suggesting a framework for integrating and analyzing data for comprehending visitor experiences at a destination.	Sentiment: SentiStrength	Results revealed three hot spots in the park where significantly pleasant messages were tweeted.
Bigne et al. (2019)	Investigating how DMOs' Twitter activity impacts hotel occupancy rate in short-break holidays.	ANN	Number of retweets and replies by users, and the number of event and tourist attraction tweets, and retweets by DMOs can predict the hotel occupancy rate.
Walker et al. (2017)	Identify determinants of obtained number of retweets in a (political) marketing context	CHAID	Retweeting occurs when the originator has a high number of followers and the sentiment is negative. Tweets with fear appeals or support for others get more retweets.

Ghiassi et al. (2013)	Developing a Twitter-specific lexicon for sentiment analysis specialized on brand-specific terms.	SVM	Researchers revealed 181 terms of sentiment expression general for Twitter and 6 brand specific ones.
Ikeda et al. (2013)	Proposing a hybrid model that blends text-based and community-based method for the demographic estimation.	SVM	By employing tweet history & follower-followee relationships, demographic information of users can be predicted.
Hawkins et al. (2016)	Using Twitter as a data stream to measure patient-perceived quality of care in US hospitals, and compare patient sentiments about hospitals.	Python library TextBlob	Twitter sentiment only weakly correlates with readmission rates but not HCAHPS ratings.
Kim et al. (2016)	Assessing competitive intelligence in social media for revealing market insight through comparing consumer opinions and sales performance of a business and one of its competitors.	SVM for purchase intention; Lexicon-based approach for sentiment	The volume of tweets points out a vital gap between the market leader and its follower. The purchase intention results also reflect this gap, but to a lesser extent. The social opinion gap is also similar to sales performance gap.

Seven studies belonging to this cluster (Hawkins et al., 2016; Kim et al., 2016; D’avanzo, 2017; Rogers et al., 2017; Walker et al., 2017; Adamopoulos et al., 2018; Park et al., 2018) conducted sentiment analysis and NLP by analyzing consumer tweets. Hawkins et al. (2016) employed over 400.000 patient tweets and conducted sentiment analysis with a Phyton library named “TextBlob” for evaluating Twitter usage as a data stream aiming to measure perceived quality of US hospitals, and found that tweet sentiment is not associated with established healthcare quality ratings. Kim et al. (2016) aimed to mine competitive intelligence for finding market insights by comparing consumer opinions and sales performance by conducting sentiment analysis with 220.000 tweets through employing SVM technique for measuring purchase intention and lexicon-based approach for sentiment analysis, and concluded that tweet volume and purchase intention discloses a noteworthy gap between the market leader and the follower. D'avanzo (2017) conducted sentiment analysis with 3.000 tweets for presenting a pipeline that enables decision makers to monitor Twitter’s sentiment on the topics of Google trends, and through employing four sentiment analysis tools (naïve Bayes detection algorithm, a simple voter algorithm, an NLTK-based sentiment analysis algorithm, commercial sentiment analysis tool) plausibility of their approach is validated. For contributing to theoretical and applied understanding of Double Jeopardy and Negative Double Jeopardy in online microblogging environment, Rogers et al. (2017) analyzed sentiment of 350.000 tweets through a statistical software package ‘Sentiment’ using Naive Bayes, and concluded that larger brands suffer from an amplified negativity amongst tweets. Walker et al. (2017) aimed anticipating number of retweets of political tweets through analyzing the sender, message sentiment, message content and structure, and through analyzing 42.000 tweets sent by British MPs during the 2015 UK General Election campaign, and accomplished it with 76,7% accuracy. Adamopoulos et al. (2018) investigated the impact of personality trait similarity on the effectiveness of WOM in Twitter by employing LDA, and found a positive relationship between the level of personality similarity and the likelihood of a subsequent purchase from a recipient of WOM message. Finally, for introducing combined utilization of Twitter geocodes, sentiment and hot spot analysis, and GIS mapping in visualization of emotions associated with on-site customer experiences, Park et al. (2018) used Sentistrength package and identified three hot spots in Disneyland where pleasant tweets were posted.

Four studies (Ghiassi et al., 2013; Ikeda et al., 2013; Pournarakis et al., 2017; Bigne et al., 2019) applied more complex machine learning methods for conducting research through employing consumer tweets, and Wang et al. (2019) by employing messages posted on micro blog Sina Weibo. Ghiassi et al. (2013) employed above a million consumer tweets for developing a sentiment classification using SVM, and by using that Twitter-specific lexicon and DAN2 machine learning approach, they found 181 terms of sentiment expressions and 6 brand specific ones with effective recall and accuracy metrics. For proposing a hybrid text and community based method for estimating demographics, Ikeda et al. (2013) employed SVM for analyzing 100.000 tweets and reached 80% accuracy rate in predicting demographic info by using tweet history and follower-user relationships. By employing SVM and LDA topic modelling methods, Pournarakis et al. (2017) devised an algorithm which is capable of generating insights on the tweet volume and topics discussed, sentiment towards a brand, optimal clustering in semantically-coherent clusters under a topic, and an overall average sentiment assessment of a given topic with 83,2% accuracy. Bigne et al. (2019) aimed assessing the impact of destination marketing organization's Twitter activity on hotel occupancy rates through employing Artificial Neural Network (ANN) method, and managed to estimate hotel occupancy rate with above 90% accuracy by using the number of retweets and replies by users, the number of event tweets, tourist attraction tweets, and retweets. In order to quantify the optimal level of marketing aggressiveness in social media for reaching the maximum popularity, Wang et al. (2019) employed Naïve Bayes method and more than 700.000 posts on Sina Weibo, and concluded that marketing aggressiveness can predict marketing popularity with 96% accuracy.

Microblogs constitute the second largest cluster of the review by involving 29% of the reviewed studies. One of the first research of the reviewed steam, which was published in 2013, belongs to this cluster. In the year 2013, research of Ghiassi et al. (2013) and Ikeda et al. (2013) gave acceleration to examination of user posts on microblogs, and Twitter is stated in the center of attention by being employed in 83% of the studies. The majority examined sentiment of the tweets for diverse purposes, and employed SVM algorithm for data analysis. Although this cluster is the second largest one, in the path leading to maturity, there is still a long way to proceed. In addition to required diverseness in Twitter-based research, other microblogging sites such as Plurk and Sina Weibo needs more attention for comprehending platform-specific dynamics.

3.1.3. Cluster 3 – Social Networking Sites

The third cluster contains "Facebook", "comment", "message", and "media type", as the main terms. Studies belonging to this cluster conducted social media analytics through analyzing data obtained from Facebook. Among six studies, four are published in business and marketing related journals, while the remaining two are in computational science related journals. The earliest date of publications belonging to this cluster is in 2013, whilst the latest ones are in 2018. Detailed information about the publications included in this cluster is provided in Table 6.

Table 6. Detailed Information Regarding the Publications in Social Networking Sites Cluster

Kwok & Yu (2013)	Revealing the types of messages that gain most clicks of “Like” and comments on Facebook.	SVM	Photo and status updates receive more likes and comments than the other two categories. Social media messages can be categorized into two message types: sales and marketing and conversational messages. Conversational messages are endorsed by more users. Media and message type impacts the number of comments a message received.
Moro et al. (2016)	Predicting the impact of posting individual messages on a social media network company's page	SVM	The type of the published content is the most relevant input feature for the model. Status updates have twice the impact of the remaining types. Posts about special offers and contests are likely to produce greater impact than product and other non-explicit brand related contents.
Lee et al. (2018)	Investigating the impact of social media advertising content on customer engagement	Combination of various different classifiers with ensemble learning	Inclusion of brand personality related attributes yield higher levels of consumer engagement. Directly informative content itself yields lower levels of engagement, but higher engagement levels are obtained when provided in combination with brand personality –related attributes.

Dhaoui et al. (2017)	Comparing performance of lexicon and machine learning approaches of automated sentiment analysis in investigation of user-generated content on social media.	ML: Investigated various algorithms but Maximum Entropy Modelling for predicting positive sentiment and the Bagging method for negative sentiment yield the best results. Sentiment: LIWC	Both approaches yield similar accuracy, and both achieve higher accuracy when classifying positive sentiment. However, they differ in classification ensembles. The combined approach yields remarkably improved performance in classifying positive sentiment.
İlhan et al. (2018)	Examine the consequences of ADA in terms of social media page volume and valence, and investigating events play role in ADA.	SVM	Negative sentiment expressed by rival brand fans make fans to defend their brand and lead to a net increase in social media performance. Main drivers: Pepsi - Advertisement, Coca Cola & Pepsi - PR, Apple & Samsung - Innovation, Apple - Announcement, Samsung - Launch
Wong et al. (2017)	Proposing an ANN-based framework for analyzing social media data in the most structured way.	ANN	It is revealed that consumers convey less information through common attributes.

Three studies (Kwok and Yu, 2013; Moro et al., 2016; Lee et al., 2018) employed Facebook posts for predicting engagement and interaction rates through machine learning. Kwok and Yu (2013) collected around 1.000 Facebook posts, employed SVM classifier for predicting number of interactions, and found that photo and status updates receive more interaction while conversational messages are being endorsed by more users. In a similar vein, by employing around 1.000 individual posts on a company's Facebook page, Moro et al. (2016) aimed predicting the impact of publishing such posts via SVM, and found that status updates result in twice the impact of the other types of posts, while special offers and contests are producing greater impact. Lee et al. (2018) employed above 106.000 Facebook messages and combination of various different classifiers with ensemble learning, and established that directly informative content results in lower levels of engagement when included in messages in isolation, but higher engagement levels may be obtained through combining it with brand personality–related attributes.

Two studies (Dhaui et al., 2017; İlhan et al., 2018) examined Facebook brand pages by employing sentiment analysis techniques. Dhaui et al. (2017) aimed evaluating and comparing performance of lexicon and machine learning approaches to automated sentiment analysis applied to UGC, and by 850 consumer comments from 83 Facebook brand pages, it is found that both approaches are similar in accuracy, but differ substantially in classification ensembles yielding a combined approach that demonstrates significantly improved performance in classifying sentiment. İlhan et al. (2018) examined the consequences of Attack Defense and Across (ADA) in terms of page volume and valence through employing 1500 comments and SVM technique, and results indicated that negative sentiment expressed by competitor brand fans encourage fans to defend their brand and causing a net increase in social media performance. On the other hand, Wong et al. (2017) aimed proposing an ANN-based approach to analyze the impact of comments on product attributes on comment length, and by employing Samsung mobile comments o Facebook, it is found that less information is conveyed through common attributes.

The third largest cluster, social networking sites, is a relatively newer one with its first publication in 2013. Acceleration in the cluster has started in 2017 and experienced its peak in 2018. The entire studies constituting the cluster are concentrated in Facebook data, and SVM is the most frequently employed algorithm. The most conspicuous aspect of the cluster is a need for diversity in data source for enhancing the progress of the field. Hence, other social networking platforms such as Tencent QQ (also known as QQ), Linkedin, and QZone requires more attention from the researchers.

3.1.4. Cluster 4 – Content Communities

The fourth cluster contains “image”, “content”, and “brand” as the main terms. Studies belonging to this cluster conducted social media analytics through employing visual and/or verbal data obtained from visual content based social media. With only three studies, this cluster is the smallest among others. Two studies are published in business and marketing related journals, while the remaining one is in a computational science related journal, and the earliest date of publication is in 2013, whilst the latest one is in 2019. Detailed information about the publications included in this cluster is provided in Table 7.

Table 7. Detailed Information Regarding the Publications in Content Communities Cluster

[redacted]	[redacted]	[redacted]	[redacted]
Jin-Jang et al. (2013)	Estimating the relationship between user profiles, value structures, and attitudes based on the replies and published content.	Euler Graph	Proposed method extracts reviewers' attitudes that have more meaning than sentiments or likes and dislikes. It also extracts utility and hedonic values as determinants of reviewers' attitudes towards the product shown on YouTube.
Klostermann et al. (2018)	Extracting brand information from social media platforms via integrating image, text and tags	N/A	Clustering Instagram posts yield three benefits; (1) Overview of what users share, (2)Allocating posts to specific context, (3) Differentiated view of brand perception
Giglio et al. (2019)	Examining the attractiveness of tourism sites by investigating the behaviour of users through social media.	Mathematica (k means), ImageIdentify	Users' behaviour identify the annual trend of photographic activity in cities.

Note: "N/A" abbreviation, which stands for not applicable, is used for the publications that did not indicate the employed algorithm and/or lexicon.

Deriving from Value Theory, Jin-Jang et al. (2013) aimed proposing a method for estimating causality between user profiles, value structures, and attitudes based on published content, and employed 2670 Youtube comments for conducting sentiment analysis. The constructed method allowed authors to extract reviewers' attitudes, two dimensional value structure (utility and hedonic) as determinants of attitude towards the product. Klostermann et al. (2018) aimed extracting brand information from Instagram by integrating image, text and tags, and hence clustered 10.000 posts according to their sentiments which provided ability of overviewing posts, allocating posts according to their content, and obtaining differentiated views on brand perception. Finally, Giglio et al. (2019) aimed determining the attractiveness of tourism sites, and by investigating 26.000 Flickr photos via k-means, ImageIdentify statistical package and clustering, it is succeeded to map the annual trend of photographic activity in cities.

Visual content cluster is the tiniest and youngest one among the others by having two out of three studies published in 2018 and 2019. Instagram, Flickr and Youtube are represented in the cluster with one study each. When the extensive usage of social media platforms centering visual content sharing, the number of studies in the field can be described as surprisingly low and restricted in terms of scope. The enduring challenge in extracting information from images and processing visual data with complicated attributes (Wang et al., 2015) may be the reason for the stated fact. However, in addition to the widespread usage of such platforms among individuals, the ongoing spread of celebrity endorsement and influencer marketing, which mainly take place on related platforms, require more attention to be directed to this field.

3.1.5. Studies Employing Cross-Media Data

Studies gathered under this heading employ cross-media data that is sourced from multiple types of social media platforms simultaneously. Among seven studies, four are published in business and marketing related journals, while the remaining are computational science related ones. The earliest date of publications belonging to this group is in 2013, whilst the latest one is in 2019. Detailed information about the publications included in this cluster is provided in Table 8.

Table 8. Detailed Information Regarding the Publications that Employ Cross-Media Data

Ryoo & Bendle (2017)	Investigating the social media strategies of 2016 U.S. presidential election candidates.	LDA Topic model	Revealed the topics on which candidates focus and showed how this changes over time.
Pineiro-Chousa et al. (2017)	Examining investors' social media activity, and social media's influence over the Chicago Board Options Exchange Market Volatility Index (VIX).	Logit, fsQCA	Social media platforms impact investors' decisions, and that this influence results in variation of the market risk. Investors' profile is vital for revealing how social network activity impacts stock markets.

Schniederjans et al. (2013)	Investigating if IM direct-assertive strategies on social media impact a firm's financial performance, and exploring the social media strategies that effect firm's financial performance.	SVM	IM direct-assertive strategies has a partial, positive impact on financial performance. Ingratiation, intimidation, organizational promotion and supplication are significantly related to financial performance.
Nguyen et al. (2018)	Proposing a framework that extracts vital sentences and posts in Web by analyzing its social context.	SVM, LDA	Local information is better in capturing summary indicators sentences and comments. Combining features from three channels helps integrating human-knowledge into summarization.
Vazquez et al. (2014)	Proposing classifiers for assigning e-wom text to one single phase of the consumer decision journey and detecting marketing mix elements.	Decision Tree	The model automatically identifies business indicators.
Liu et al. (2016)	Combining diverse methods (cloud computing, machine learning, text mining) for providing a framework to employ online content for forecasting demand.	Dynamic panel data linear model, Information Measures of Content2Count, Sentiment, and n-grams PCA	Tweets, Google searches, Wikipedia views, IMDB reviews, and Huffington Post News volume have a positive impact on TV ratings. Carefully summarized Tweet content has the highest predictive power.
Pelaez et al. (2019)	Developing a decision-making model that creates alternative ranking by contextualizing unsolicited opinions expressed by decision-makers in social media.	N/A	The proposed model yielded an alternative ranking that contextualizes the feelings/opinions represented through feeling consistent with data extracted from social media.

Note: “N/A” abbreviation, which stands for not applicable, is used for the publications that did not indicate the employed algorithm and/or lexicon.

Ryoo and Bendle (2017) and Pineiro-Chousa et al. (2017) conducted social media analytics by obtaining data through a microblog (Twitter) and a social networking site (Facebook). Ryoo and Bendle (2017) examined social media strategies of presidential election candidates by analyzing 85.000 Twitter and Facebook posts through LDA Topic Model and exposed the topics that the candidates focus on over time. Pineiro-Chousa et al. (2017) analyzed the impact of investors’ Twitter and Facebook activities over the Chicago Board Options Exchange Market Volatility Index (VIX) by analyzing over 90.000 investor posts through Logit and fsQCA, and the results indicate that combinations among different social network variables effect decisions of investors, which, in turn, leads to a variation of the market risk.

Vazquez et al. (2014) and Liu et al. (2016) employed Natural Language Processing (NLP) method for analyzing social media data obtained from diverse networks. Vazquez et al. (2014) aimed assigning eWOM texts to a single phase of the consumer decision excursion and identifying marketing mix elements by analyzing around 23.000 posts on various social media platforms through decision tree, and achieved 78% accuracy in identifying business indicators automatically. Liu et al. (2016) aimed estimating content preferences by applying HDLDA to users’ search queries, and their application successfully managed to predict CTR in sponsored search advertising.

Among remaining three studies, Schniederjans et al. (2013) and Nguyen et al. (2018) employed SVM method while Pelaez et al. (2019) applying ISMA-OWA operator for analyzing social media data. Schniederjans et al. (2013) investigated the extent to which impression management (IM) direct-assertive strategies in social media impact a firm's financial performance by analyzing social media data of 150 publicly traded firms through SVM and concluded that ingratiation, intimidation, organizational promotion and supplication strategies can estimate financial performance with 75,25% accuracy rate. Nguyen et al. (2018) aimed automatically mining key sentences and posts of a Web document by integrating its social context, and by analyzing UGC through SVM and LDA, it is found that social context supports local information in capturing summary indicators in sentences. Finally, Pelaez et al. (2019) developed a decision-making model to achieve an alternate ranking that contextualizes opinions expressed in social media by employing 60.000 social media comments through ISMA-OWA method.

The cross-fertilization of diverse social media platforms has started in 2013 and gained a remarkable momentum. When the methodological structure of these studies are examined, it can be noticed that the diversity in data sources require more complicated algorithms for analysis.

However, it should also be highlighted that the novelty in research designs and data analysis produce accordingly prolific outcomes. As the progress in the research area matures, the number of such studies are expected to increase with the intention of having a more holistic view and broader understanding. Moreover, since both individuals and companies have presence in multiple social media platforms, employing cross-media data would yield more accurate results for researchers and practitioners.

4. Summary and Concluding Remarks

SMA is an emerging domain with various applications of machine learning algorithms for processing high-volume and complex data in a cost effective way, deriving insights from user generated content. This study aimed presenting an integrative framework by portraying machine learning application trends and approaches in SMA, and revealed five distinct research clusters. Findings point out that application of machine learning algorithms has gained a momentum and growing diversely in SMA domain. When the multifaceted nature of the research area is considered, it may be stated that numerous undiscovered or under-analyzed paths exist for researchers to be recognized and examined. Therefore, for the near future, application of machine learning algorithms to the domain is expected to continue growing exponentially as the extensive usage of social media, and the bidirectional interaction of companies and consumers through social media platforms endure.

For benefitting from the multifaceted and fruitful nature of the domain, and broadening the research agenda, it is vital to gather scholars from diverse academic backgrounds and practitioners. Such an attempt plays a pivotal role in advancement of the field in two ways. First, obtaining opinions and suggestions of practitioners and companies may benefit strengthening the bridge between academia and practice while highlighting the needs of companies and pointing new research avenues. Second, scholars with marketing, management, and business background may bring new theoretical and contextual perspectives to the domain, and researchers with computational science background may solidify such attempts with novel methodological perspectives and tools that are not widely applied in social sciences domain. Hence, cross-fertilization may enrich the maturity level of the domain by enhancing depth and width of extant body of knowledge.

This study contributes to social media analytics and machine learning domains in three ways. First, state-of-art analysis highlights interdisciplinary and multifaceted nature of social media analytics through plotting the extant research and the interrelations between two research domains. Second, this article maps various intellectual standpoints and intersections in the mentioned fields and augments the research mapping with interpretation of the consequential research clusters. Lastly, through defining research clusters, this study provides an innovative topical classification and a guideline on this evolving research domain indicating different research streams to both prevent inefficiency caused by studies with iterative scope and design, and to associate two diverse domains to reach novel insights.

5. Limitations and Directions for Future Research

Even though this review is one of the first attempts to systematically map the field, it contains its own limitations. It should be noted that findings of this study should be remarked within the context of five main limitations: (1) The data collection was conducted using solely the *ISI Web of Knowledge* database, and therefore not all journals and issues are covered; (2) For ensuring consistency of the sample, research published in refereed international journals only were considered, and publications in other forms, such as books, conference proceedings, and research reports were not included; (3) Articles in “forthcoming” status, and conceptual studies were not included; (4) Reviewed studies were acquired from English-language journals; (5) Even though every exertion was made to leave no avenue unexplored by collecting and reviewing all relevant research in the domain, the risk of overlooking relevant studies remains.

In line with the findings, two main avenues for future research is suggested. First, giving countenance to comprehensive literature reviews would play a vital role for banding research efforts in diverse domains together and keeping a close watch on the improvements in the field during the expansion period. Such efforts may benefit the research area by detecting current gaps and expediting the maturity level. Moreover, presence of a wider range of publications including books, proceedings, and journals in other databases that are excluded in this study may yield wider variety of studies, and hence may provide a more holistic view on the subject. Second, as the literature broadens, bibliometric techniques may provide more fruitful insights on the hidden relationship patterns such as the concealed scholarly communities, which can be discovered by citation, co-citation and bibliographic coupling analyses (Van Eck and Waltman, 2014).

REFERENCES

Adamopoulos, P., Ghose, A., & Todri, V. (2018). The impact of user personality traits on word of mouth: Text-mining social media platforms. *Information Systems Research*, *29*(3), 612-640.

Agnihotri, A., & Bhattacharya, S. (2016). Online review helpfulness: Role of qualitative factors. *Psychology & Marketing*, *33*(11), 1006-1017.

Ali, F., Kwak, K. S., & Kim, Y. G. (2016). Opinion mining based on fuzzy domain ontology and Support Vector Machine: A proposal to automate online review classification. *Applied Soft Computing*, *47*, 235-250.

Beyer, M. A., & Laney, D. (2012). The importance of 'big data': a definition. *Stamford, CT: Gartner*, 2014-2018.

Bigne, E., Oltra, E., & Andreu, L. (2019). Harnessing stakeholder input on Twitter: A case study of short breaks in Spanish tourist cities. *Tourism Management*, *71*, 490-503.

Bilro, R. G., Loureiro, S. M. C., & Guerreiro, J. (2019). Exploring online customer engagement with hospitality products and its relationship with involvement, emotional states, experience and brand advocacy. *Journal of Hospitality Marketing & Management*, *28*(2), 147-171.

Calheiros, A. C., Moro, S., & Rita, P. (2017). Sentiment classification of consumer-generated online reviews using topic modeling. *Journal of Hospitality Marketing & Management*, *26*(7), 675-693.

Chaffey, D. (2019). Global Social Media Research Summary 2019. Available at: https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/ Accessed 10 July 2019.

Costa, A., Guerreiro, J., Moro, S., & Henriques, R. (2019). Unfolding the characteristics of incentivized online reviews. *Journal of Retailing and Consumer Services*, *47*, 272-281.

D'Avanzo, E., Pilato, G., & Lytras, M. (2017). Using Twitter sentiment and emotions analysis of Google Trends for decisions making. *Program*, *51*(3), 322-350.

DEMİRCİ, S., & SAĞIROĞLU, Ş. (2017). Sosyal Ağ Verilerinin Kullanım Alanları Üzerine Kapsamlı Bir İnceleme. *Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji*, *5*(2), 1-21.

Dhaoui, C., Webster, C. M., & Tan, L. P. (2017). Social media sentiment analysis: lexicon versus machine learning. *Journal of Consumer Marketing*, *34*(6), 480-488.

Eslami, S. P., & Ghasemaghaei, M. (2018). Effects of online review positiveness and review score inconsistency on sales: A comparison by product involvement. *Journal of Retailing and Consumer Services*, *45*, 74-80.

Fabbri, S., Hernandes, E., Di Thommazo, A., Belgamo, A., Zamboni, A., & Silva, C. (2012, June). Using information visualization and text mining to facilitate the conduction of systematic literature reviews. In *International Conference on Enterprise Information Systems*, pp. 243-256. Springer, Berlin, Heidelberg.

Felizardo, K. R., Salleh, N., Martins, R. M., Mendes, E., MacDonell, S. G., & Maldonado, J. C. (2011, September). Using visual text mining to support the study selection activity in systematic literature reviews. In *2011 International Symposium on Empirical Software Engineering and Measurement*, pp. 77-86.

Ghiassi, M., Skinner, J., & Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with applications*, *40*(16), 6266-6282.

Giglio, S., Bertacchini, F., Bilotta, E., & Pantano, P. (2019). Using social media to identify tourism attractiveness in six Italian cities. *Tourism Management*, *72*, 306-312.

Greenwood, B. N., & Gopal, A. (2015). Research note—Tigerblood: Newspapers, blogs, and the founding of information technology firms. Information Systems Research, 26(4), 812–828.

Gurzki, H., & Woisetschläger, D. M. (2017). Mapping the luxury research landscape: A bibliometric citation analysis. *Journal of Business Research*, *77*, 147-166.

Harzing, A. W. (2013). Document categories in the ISI Web of Knowledge: Misunderstanding the social sciences?. *Scientometrics*, *94*(1), 23-34.

Hawkins, J. B., Brownstein, J. S., Tuli, G., Runels, T., Broecker, K., Nsoesie, E. O., ... & Greaves, F. (2016). Measuring patient-perceived quality of care in US hospitals using Twitter. *BMJ Qual Saf*, *25*(6), 404-413.

Heng, Y., Gao, Z., Jiang, Y., & Chen, X. (2018). Exploring hidden factors behind online food shopping from Amazon reviews: A topic mining approach. *Journal of Retailing and Consumer Services*, *42*, 161-168.

Hu, Y. H., Shiau, W. M., Shih, S. P., & Chen, C. J. (2018). Considering online consumer reviews to predict movie box-office performance between the years 2009 and 2014 in the US. *The Electronic Library*, *36*(6), 1010-1026.

Ikeda, K., Hattori, G., Ono, C., Asoh, H., & Higashino, T. (2013). Twitter user profiling based on text and community mining for market analysis. *Knowledge-Based Systems*, *51*, 35-47.

Ilhan, B. E., Kübler, R. V., & Pauwels, K. H. (2018). Battle of the brand fans: impact of brand attack and defense on social media. *Journal of Interactive Marketing*, *43*, 33-51.

Jang, H. J., Sim, J., Lee, Y., & Kwon, O. (2013). Deep sentiment analysis: Mining the causality between personality-value-attitude for analyzing business ads in social media. *Expert Systems with applications*, *40*(18), 7492-7503.

Jimenez-Marquez, J. L., Gonzalez-Carrasco, I., Lopez-Cuadrado, J. L., & Ruiz-Mezcua, B. (2019). Towards a big data framework for analyzing social media content. *International Journal of Information Management*, *44*, 1-12.

Kapoor, K.K., Tamilmani, K., Rana, N.P., Patil, P., Dwivedi, Y.K., Nerur, S., 2017. Advances in social media research: past, present and future. Inform. Syst. Front.

Kim, Y., Dwivedi, R., Zhang, J., & Jeong, S. R. (2016). Competitive intelligence in social media Twitter: iPhone 6 vs. Galaxy S5. *Online Information Review*, *40*(1), 42-61.

Klostermann, J., Plumeyer, A., Böger, D., & Decker, R. (2018). Extracting brand information from social networks: Integrating image, text, and social tagging data. *International Journal of Research in Marketing*, *35*(4), 538-556.

Kwok, L., & Yu, B. (2013). Spreading social media messages on Facebook: An analysis of restaurant business-to-consumer communications. *Cornell Hospitality Quarterly*, *54*(1), 84-94.

Lee, D., Hosanagar, K., & Nair, H. S. (2018). Advertising content and consumer engagement on social media: evidence from Facebook. *Management Science*, *64*(11), 5105-5131.

Liu, X., Singh, P. V., & Srinivasan, K. (2016). A structured analysis of unstructured big data by leveraging cloud computing. *Marketing Science*, *35*(3), 363-388.

Meredith, J. (1993), "Theory building through conceptual methods", *International Journal of Operations & Production Management*, Vol. 13 No. 5, pp. 3-11.

Mergel, G. D., Silveira, M. S., & da Silva, T. S. (2015, April). A method to support search string building in systematic literature reviews through visual text mining. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, pp. 1594-1601.

Moro, S., Rita, P., & Vala, B. (2016). Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. *Journal of Business Research*, *69*(9), 3341-3351.

Nave, M., Rita, P., & Guerreiro, J. (2018). A decision support system framework to track consumer sentiments in social media. *Journal of Hospitality Marketing & Management*, *27*(6), 693-710.

Nguyen, M. T., Tran, D. V., & Nguyen, L. M. (2018). Social context summarization using user-generated content and third-party sources. *Knowledge-Based Systems*, *144*, 51-64.

Nilashi, M., Ibrahim, O., Yadegaridehkordi, E., Samad, S., Akbari, E., & Alizadeh, A. (2018). Travelers decision making using online review in social network sites: A case on TripAdvisor. *Journal of computational science*, *28*, 168-179.

Park, S. B., Kim, H. J., & Ok, C. M. (2018). Linking emotion and place on Twitter at Disneyland. *Journal of Travel & Tourism Marketing*, *35*(5), 664-677.

Peláez, J. I., Martínez, E. A., & Vargas, L. G. (2019). Decision making in social media with consistent data. *Knowledge-Based Systems*, *172*, 33-41.

Piñeiro‐ Chousa, J., Vizcaíno‐ González, M., & Pérez‐ Pico, A. M. (2017). Influence of social media over the stock market. *Psychology & Marketing*, *34*(1), 101-108.

Porter, A., Kongthon, A., & Lu, J. C. (2002). Research profiling: Improving the literature review. *Scientometrics*, *53*(3), 351-370.

Pournarakis, D. E., Sotiropoulos, D. N., & Giaglis, G. M. (2017). A computational model for mining consumer perceptions in social media. *Decision Support Systems*, *93*, 98-110.

Randhawa, K., Wilden, R., & Hohberger, J. (2016). A bibliometric review of open innovation: Setting a research agenda. *Journal of Product Innovation Management*, *33*(6), 750-772.

Rathan, M., Hulipalled, V. R., Venugopal, K. R., & Patnaik, L. M. (2018). Consumer insight mining: Aspect based twitter opinion mining of mobile phone reviews. *Applied Soft Computing*, *68*, 765-773.

Rogers, A., Daunt, K. L., Morgan, P., & Beynon, M. (2017). Examining the existence of double jeopardy and negative double jeopardy within Twitter. *European Journal of Marketing*, *51*(7/8), 1224-1247.

Ryoo, J., & Bendle, N. (2017). Understanding the social media strategies of US primary candidates. *Journal of Political Marketing*, *16*(3-4), 244-266.

Schneier, B. (2010). A taxonomy of social networking data. *IEEE Security & Privacy*, *8*(4), 88-88.

Schniederjans, D., Cao, E. S., & Schniederjans, M. (2013). Enhancing financial performance with social media: An impression management perspective. *Decision Support Systems*, *55*(4), 911-918.

Shareef, M. A., Mukerji, B., Alryalat, M. A. A., Wright, A., & Dwivedi, Y. K. (2018). Advertisements on Facebook: Identifying the persuasive elements in the development of positive attitudes in consumers. *Journal of Retailing and Consumer Services*, *43*, 258-268.

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics–Challenges in topic discovery, data collection, and data preparation. *International journal of information management*, *39*, 156-168.

Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, *57*, 117-126.

Van Eck, N. J., & Waltman, L. (2014). Visualizing bibliometric networks. In Y. Ding, R. Rousseau, & D. Wolfr

Vázquez, S., Muñoz-García, Ó., Campanella, I., Poch, M., Fisas, B., Bel, N., & Andreu, G. (2014). A classification of user-generated content into consumer decision journey stages. *Neural Networks*, *58*, 68-81.

Walker, L., Baines, P. R., Dimitriu, R., & Macdonald, E. K. (2017). Antecedents of retweeting in a (political) marketing context. *Psychology & Marketing*, *34*(3), 275-293.

Wang, X., Baesens, B., & Zhu, Z. (2019). On the optimal marketing aggressiveness level of C2C sellers in social media: Evidence from china. *Omega*, *85*, 83-93.

Wang, Y.,Wang, S., Tang, J., Liu, H., & Li, B. (2015). Unsupervised sentiment analysis for social media images (pp. 2378–2379). Proceedings of the 24th International Joint Conference on Artificial Intelligence.

Wong, T. C., Chan, H. K., & Lacka, E. (2017). An ANN-based approach of interpreting user-generated comments from social media. *Applied Soft Computing*, *52*, 1169-1180.

Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management, 58*, 51-65.