POST-PRINT VERSION

UNRAVELING e-WOM PATTERNS USING TEXT MINING AND SENTIMENT ANALYSIS

João Guerreiro, Sandra Loureiro
Instituto Universitário de Lisboa (ISCTE-IUL), Portugal

ABSTRACT

Electronic word-of-mouth (e-WOM) is a very important way for firms to measure the pulse of its online reputation. Today, consumers use e-WOM as a way to interact with companies and share not only their satisfaction with the experience, but also their discontent. E-WOM is even a good way for companies to co-create better experiences that meet consumer needs. However, not many companies are using such unstructured information as a valuable resource to help in decision making. First, because e-WOM is mainly textual information that needs special data treatment and second, because it is spread in many different platforms and occurs in near-real-time, which makes it hard to handle. The current chapter revises the main methodologies used successfully to unravel hidden patterns in e-WOM in order to help decision makers to use such information to better align their companies with the consumer's needs.

Keywords: e-WOM, Text Mining, Sentiment Analysis, NLP, LDA, CTM.

INTRODUCTION

Today, e-WOM is an extremely important source for Marketing due to its impact on the online reputation of the firms. Consumers are no longer passive bystanders. Following consumer satisfaction expressed through online interactions on Facebook, Twitter, Instagram and other user generated content sites is paramount for effectively implementation of corrective measures that may increase satisfaction and consumer engagement with the brands (Bilro, Loureiro & Guerreiro, 2018).

However, although companies have been digitalizing themselves and upgrading their infrastructure to accommodate such Big Data with technology that grabs all the interactions with the consumer in real-time (written or verbal), there is still a long work to do regarding the effective use of such information to unravel hidden patterns of behavior. However, there are some successful examples of using such information to help decision-making. In Tourism, companies such as ReviewPro and Revinate offer complete solutions for firms to grasp the unstructured information written in sites such as TripAdvisor and Booking about their brands and their competitors. Using such information, companies may understand how their online reputation is changing over time and improve guest experience according to their feedback (Nave, Rita & Guerreiro, 2018). However, such information is often offered as a silo of information and not integrated with the company's' remaining key performance indicators (KPIs). To do so, companies must integrate analytical skills and develop internal decision support systems that may able them to integrate both structured (e.g. Financial KPIs, Human Resources KPIs), and unstructured information (e.g. reviews, online posts on the company's Facebook page, verbal complains).

The current chapter analyzes the characteristics of e-WOM and presents a theoretical approach to the most relevant methods used to handle unstructured data. Such information may allow managers to treat e-WOM data in order to uncover hidden patterns of behavior.

BACKGROUND

The emergence of the Web 2.0 has brought a new era of consumer-brand interaction through the spread of electronic word-of-mouth. E-WOM may be defined as "all informal communications directed at consumers through Internet-based technology related to the usage or characteristics of particular goods and services, or their sellers." (Litvin, Goldsmith & Pan, 2008, pp.461).

While in the early days of the Internet, companies had mainly a one way communication with their consumers through institutional web sites, today users interact with companies in a two-way communication. The consumer today is both the listener and the originator of information and such change echoed for the entire decision-making process. Not only in the awareness of need stage, where consumers may interact with viral communication videos and write their opinion or share such communication with their network of friends, but also while searching for alternatives online, where consumers read and form an opinion about the experiences or products in the market, or in the purchase and post-purchase stage, where some consumers are even driven to express their own opinion about the experience. Motivations of such behavior vary from (1) a need to have a platform to spread a message for an assistance, (2) to share negative feelings, (3) by a genuine concern toward other users, (4) for extraversion and self-enhancement, (5) for social and economic benefits, (6) to help the brand or (7) to seek for advices (Hennig-Thurau et al., 2004). Some of them are positive drivers and may help the companies to achieve a better reputation online, but some are negative drivers that may harm the company if not properly addressed.

Companies have been trying to keep up with such progress by (1) setting specialized teams of digital marketers responsible for handling such interaction and (2) investing on Big Data infrastructure that captures all this information in near-real-time for later analysis. E-WOM is usually posted online in the form of textual messages either in social media or in recommendation sites. However, today bloggers also share e-WOM through video, and that information may also have valuable information for brands to understand how are they being viewed and discussed online. Therefore, all public information spread online (text, audio, video) should be captured in Big Data systems (usually also transformed into a single type of media such as text) for helping brands to better align their positioning with the expectations of their consumers.

Despite the recent technological evolution in Big Data infrastructure, allowing information with such volume, variety and velocity to be captured and stored efficiently, there is still a need to analyze information and transform it into useful patterns that may be helpful for decision-making. Text Mining (TM) has been used (along with Natural Language Processing techniques and Sentiment Analysis) to successfully grasp the hidden patterns in data and present the most relevant drivers of behavior stemming from e-WOM data.

TEXT MINING

Sanchez et al. (2008) define text mining as the discovery of non-trivial, previously unknown and potentially useful information from text. TM is a form of semi-structured analysis of unstructured data that dates back to the work of Hearst (1999). Although unstructured data has been around since companies started to keep textual documents in their database systems, only recently, technology allowed the huge amount of information stored in such systems to be thoroughly analyzed. Today, Big Data infrastructures allow companies to gather not only documents but also real-time textual information such as tweets, posts, complains over the call-center or any other type of interaction with the consumer. There are generally two types of textual analysis. The first is a deductive approach which uses a top-down approach following pre-determined associations and relations between words. Such words are included in an ontology or dictionary that determines much of the process of knowledge discovery on data (Hristovski, Peterlin, Mitchel and Humphrey, 2005). The alternative type is a bottom-up approach, were unstructured data is structured into a set of terms that are then

used to uncover latent relationships in text or classify specific events (such as for example the event of a fake news or a reputational issue) using machine learning algorithms. Therefore, the second approach combines the use of text mining as a way to structure data and then uses traditional data mining techniques to uncover patterns in the data (Sanchez et al., 2008).

Regarding the use of TM in e-WOM, usually both approaches are combined, particularly because consumers may write anything that comes to their mind and sometimes a formal dictionary may help the analysis on a big collection of data to focus on the most important elements for a specific sector.

In the inductive approach the work starts by extracting the data to a workable set of documents or *corpus*. Usually in e-WOM analysis, each review, tweet or post defines the *corpus* that together sets up the *corpora* (Feinerer, Hornik & Meyer, 2008). After the initial stage of data collection a preparation stage follows. In some cases, e-WOM may be extracted in real-time to feed the next stages.

The preparation stage converts the *corpus* into a set of bag-of-words (a group of relevant terms) for analysis. However, in order to structure the text into relevant terms, its semantic context has to be taken into consideration.

Natural Language Processing

Natural Language Processing (NLP) are a set of techniques that capture the semantic characteristics of text so that later analysis may take such context into consideration. NLP tasks include tokenization, part-of-speech tagging or named entity recognition (Collobert et al., 2011).

Tokenization breaks the text into small tokens that may be single terms or n-terms depending on common relations and context. For example, depending on the semantic context of a review, the expression “alarm clock” may have different meaning if they are assumed to be a single token or two separate tokens “alarm” and “clock”. A proper tokenization is then an extremely important part of the deductive analysis of text (Hassler & Fliedl, 2006). Another NLP technique that helps on defining contextual meaning is the part-of-speech (POS) task. POS classifies text tokens according to its syntactic role (noun, verb, adverb). Depending on the context of a review, the same word may have different syntactic roles. For example, the bi-term “fast” expressed in a review may be used as an adjective: “This mobile phone is fast” or as an adverb: “The mobile phone is loosing battery fast”. Therefore, a POS transformation technique ensures that when grouping text, only those tokens that are common are grouped to form a relationship between tokens (e.g. words) and documents (e.g. reviews). A final preparation step usually rips text from its *stopwords* (the set of auxiliary terms that are not relevant for analysis after semantic classification is performed). Punctuations and words such as “a”, “he”, “she”, “for” may be removed from text for building the document-term-matrix. In many situations, some techniques to reduce complexity may also be applied such as stemming and lemmatization. Stemming is a heuristic method that reduces each term to its radical term (*stem*) so that words with the same radical term may be analyzed together (Porter, 1980). For example, words such as “run” and “running” may be analyzed together after a stemming procedure. On the other hand, a lemmatization approach takes into consideration vocabulary and morphology and is therefore a more advanced technique.

Document-Term-Matrix and Wordclouds

A first exploratory analysis of text after transformation may be done using the document-term-matrix (DTM) analysis. As the name implies, the DTM is a cross-relation between each document (e.g. a review) and each token. Although a sum approach may be used to fill such matrix (term frequency), the term-frequency-inverse document frequency (TF-IDF) is usually the best approach to reduce the sparsity of the matrix (composed of many zero values on the crossing between reviews and terms) (Grün & Hornik, 2011). The TF-IDF approach weights differently the terms in each document so that the terms that occur more often in a single document but not often in all the documents are more relevant (Delen & Crossland, 2008).

The DTM may be explored using a *wordcloud*, a graphical representation of the weight of each term either by using the absolute frequency or by using the TF-IDF.

Clustering of e-WOM

After an exploratory analysis, the text may be grouped into clusters of words. Although traditional clustering techniques such as k-Means may be used to group text into different groups, a more appropriate approach is the use of mixed-membership clustering techniques such as topic models. Topic models are "probabilistic models for uncovering the underlying semantic structure of a document collection based on a hierarchical Bayesian analysis of the original texts" (Blei & Lafferty, 2007, p. 1). They can be represented as a relationship between an observable *N* number of words in a *D* document and a latent set of *K* variables. A topic (*t*) is modelled as a multinomial probability distribution over a set of words (*w*) in a given document, such as $p(w|t)$, for $t \in 1{:}K$ (Blei, Ng, & Jordan, 2003).

Each latent topic is a distribution over words in the document where each term has a different probability to belong to that underlying topic. Also, they are mixed-membership models, given that each document can belong simultaneous to multiple topics at the same time (Grün & Hornik, 2011). Topics are hidden entities that can be inferred from observable words using posterior inference. By analyzing topics instead of a bag-of-words, text mining can find useful structure in the *corpus* collection of e-WOM information (Blei & Lafferty, 2007). Two of the most commonly used topic model algorithms are Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) and Correlated Topic Models (CTM) (Blei & Lafferty, 2007).

Latent Dirichlet Allocation

In Latent Dirichlet Allocation, the assumptions relies in a generative process that it is believed to have been used to produce the *corpus* (Blei, Ng, & Jordan, 2003).The algorithm is based on Latent Semantic Indexing and probabilistic Latent Semantic Indexing (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990; Hofmann, 1999), and assumes that *corpora* are written as follows:

a) First, a decision is made about the number of topics the document will have;
b) A second step defines the proportions of each topic in the document, i.e. for example in a review about the service in a restaurant, 80% may be related to staff and 20% may be related to the price;
c) Afterward, words are given proportions according to their importance for each topic;
d) A word is taken according to its importance in the topic, and according to the topic distribution.

Figure 1 shows an example using the abstract of a paper from Strahilevitz & Myers (1998), in which, the histogram to the right represents the topic distribution over the documents that the generative process infers from the text.

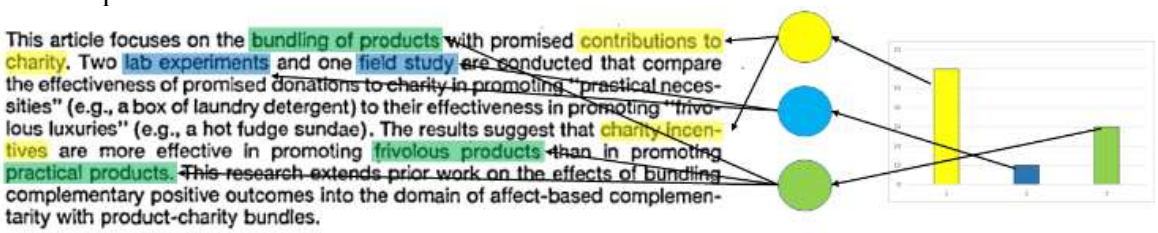

Figure 1. Topic Models Generative Process

Although this generative process only generates a meaningless bag-of-words, it is useful for the LDA algorithm purposes, which is, to generate a stochastic process that represents the hidden model and then to reverse it using posterior probabilities.

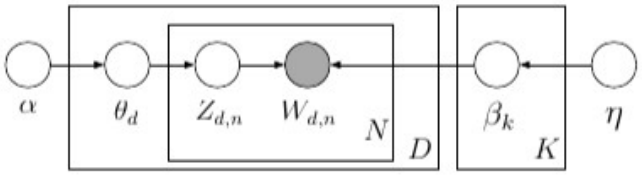

Figure 2. LDA generation procedure (Blei and Lafferty, 2009)

Figure 2 shows the graphical model representation of LDA generation procedure which is described by Blei & Lafferty (2009) as:

(1) For each topic *K*,
   a. Draw a distribution over words, $\vec{\beta}_k\ Dir_y(\eta)$.
(2) For each document *D*,
   a. Draw a vector of topic proportions, $\vec{\theta}_d \sim Dir_y(\vec{\alpha})$.
   b. For each word,
      i. Draw a topic assignment $Z_{d,n} \sim Mult(\vec{\theta}_d), Z_{d,n} \in \{1 \ldots . K\}$.
      ii. Draw a word $W_{d,n} \sim Mult(\vec{\beta} z_{d,n}), W_{d,n} \in \{1 \ldots . V\}$.

V is the size of the vocabulary of the corpus document, $\vec{\alpha}$ a positive K-vector and η a scalar.

As the name implies, LDA algorithm uses Dirichlet distribution to design the generative process. While $Dir_y(\vec{\alpha})$ is a V-dimensional Dirichlet, $Dir_y(\eta)$ is the distribution of the distributions using a dimensional symmetric Dirichlet (Blei & Lafferty, 2009). A Dirichlet distribution is used because their properties ensure that the distribution of a Dirichlet distribution is still a Dirichlet distribution with the same characteristics, which is useful for computational purposes. However, the inference over the stochastic process of LDA has some shortcomings. The posterior distribution of the hidden variables is computationally intractable. The effort needed to compute the integral in the normalized distribution used in LDA is NP-hard and must be approximated using an approximate inference technique (Blei & Lafferty, 2009). Multiple techniques have been used to optimize this equation, such as collapsed variational inference (Teh et al., 2006), expectation propagation (Minka & Lafferty, 2002), and Gibbs sampling (Griffiths & Steyvers, 2007).

Correlated Topic Models

Although LDA has been successfully applied to model latent topics in documents, they lack an important characteristic. The Dirichlet distribution assumes that its vector points are nearly independent, which means that when topics are modelled using the Dirichlet distribution, they are assumed to be independent (Blei & Lafferty, 2009). However, in a real-word review, topics are usually correlated. Correlated Topic Models (CTM) builds on LDA but modifies the distribution used to model the topic proportions. Instead of using the Dirichlet distribution, CTM uses a logistic normal distribution (Atchison & Shen, 1980). The logistic normal distribution incorporates the covariance among the topics.

SENTIMENT ANALYSIS

Sentiment analysis goes a step further than identifying the more relevant terms or grouping text into topics. The main purpose of such task is to identify the polarity (or a sentiment scale) of the corpus, the sentences of the corpus or even individual n-grams. Sentiment analysis is a crucial step to extract relevant patterns in eWOM. Other than just knowing what consumers are discussing, managers want to know if they are discussing it in a more positive or negative tone. Such information allows

managers to focus on corrective measures to address the problems identified and therefore to increase customer satisfaction.
Sentiment analysis may be developed using a machine learning approach or a lexicon based approach (Medhat, Hassan, & Korashy, 2014). The machine learning approach is usually used to identify polarities (or emotions such as anger, disgust, fear, interest, surprise, etc.) in entire reviews or sentences. Using a training dataset with reviews or sentences already classified with the different emotions, machine learning algorithms such as Artificial Neural Networks (ANN) or Support Vector Machines (SVM) create a model that may be used to predict emotions on a validation dataset. If the model is accurate enough it can be used to predict emotions on a new set of reviews. The second approach (lexicon based) is based on a list of words (a seed list) that contains the word polarity. Such seed lists may be created using a dictionary based approach or a corpus based approach (Feldman, 2013). The dictionary based approach is a top down method that is usually created using a set of starting words and their polarities, which are then expanded through the use of synonyms and antonyms. Wordnet is an example of a lexical database that contains more than 117.000 words and its synonym relations and is often used to expand initial seed of words in sentiment analysis (Fellbaum, Christiane, 2005). Although a bottom-up approach may be used successfully to uncover sentiments in text, it often lacks the specific terms of the different business domains. For example the word "short" may have different sentiment polarities if we are discussing the time spent to serve a customer on a restaurant – positive sentiment, or if it is referring to the size of a shirt – negative sentiment. Therefore, the dictionary based approach is often coupled with the *corpus* based approach that uses the text itself to identify word polarities. The *corpus* based approach also depends on seed lists but then uses them with natural language processing (NLP) techniques to classify the text (Caro & Grella, 2013; Liu & Zhang, 2012). There are several lexicons that are available online to be used for sentiment analysis purposes such as Hamilton, Clark, Leskovec & Jurafsky (2016) domain-specific sentiment lexicon and the AFFIN (Nielsen, 2019), BING (Bing, 2019) and NRC (Mohammad, 2019) lexicons available on the *tidytext* package on R (a statistical tool often used for text mining and sentiment analysis). While BING has a binary classification of positive and negative words, AFFIN has a set of words with sentiments classified between -5 (more negative) to 5 (more positive), and NRC classifies a set of words with ten categories (positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise and trust).

Issues, Controversies, Problems

Despite the innumerous advantages of using semi-automatic techniques to extract patterns from text, there are still some challenges to overcome in text mining and sentiment analysis. Although there are multiple standard lexicons that may be used in most situations, there are several specific industries were the same word may have a different meaning or a different polarity. As highlighted above, for example, the word "fast" may be a positive thing if the consumer is talking about the waiting time in a restaurant, but a negative thing if the consumer is discussing "fast" food. Also, although NLP has evolved tremendously in the last decade (especially in handling the English language) there are still several issues regarding the understanding of morphologically rich languages (e.g. Arabian, Hebrew) (Kincl, Novák & Přibil, 2019). The identification of irony, sarcasm and humour also still presents a big challenge for scholars in the field (Katyayan & Joshi, 2019; Farias & Rosso, 2017).

SOLUTIONS AND RECOMMENDATIONS

Solutions for the problems presented above include word sense disambiguation, in which machine learning techniques are used to classify terms depending on their context (Vechtomova, 2017). In fact, machine learning techniques have also been used to detect irony and sarcasm. Using pre-classified expressions of irony or sarcasm it is possible to predict new sentences by using algorithms such as support vector machines (SVM), artificial neural networks (ANN) and others. For example, Bharti et al., (2017) used SVM and Naïve Bayes classifiers, while Mukherjee and Bala (2017) used Naïve Bayes and fuzzy clustering to address such issues.

FUTURE RESEARCH DIRECTIONS

The current paper addresses the main methods that have been used to highlight behavior patterns in e-WOM messages. However, there are new approaches that have been recently suggested by scholars and that complement the vision here presented, such as the use of the network relations between the several consumers that share e-WOM messages. The connections between consumers in a network (called a graph) may be important to determine if for example a negative opinion may spread through relations and at what speed. To study such problem, graph mining techniques have been employed recently to detect consumer communities, opinion leaders and to study network dynamic. A good review may be found on the work of Bamakan, Nurgaliev and Qu, 2019, where the authors describe a methodological review of the use of graph mining to handle consumer interactions and how some consumers may lead others in their opinion, making them important actors in the network. A positive opinion leader may become an evangelist of the brand and should therefore be incentivized to share its motivation with its peers, while a negative opinion leader may harm the company's' reputation and firms should be particularly careful when addressing its needs (Bamakan, Nurgaliev & Qu, 2019; Arrami, Oueslati & Akaichi, 2017; Bilici & Saygın, 2017; Chen et al., 2017).

Another new promising direction to handle e-WOM is to use adaptive techniques such as deep learning algorithms to handle unstructured data (Arora & Kansal, 2019). Deep learning algorithms have recently gained traction to support self-driving cars and adapt easily to new variables in the environment (in this case, new words and expressions in e-WOM). Therefore, it is a promising direction to increase e-WOM classification accuracy in the future.

CONCLUSION

e-WOM text starts as a set of unstructured review, comment, post, in the form of free text. However, the value in using such information for understanding consumer behavior lies in trying to make sense on the patterns of data that rely latent in the comments of multiple consumers at the same time. To perform such task, scholars have developed a set of techniques based on natural language processing (NLP) that analyses the semantic relations of the text and structures the most important terms discussed by consumers. A second step aggregates the relevant terms into latent topics using techniques such topic models. Finally, sentiment analysis classifies such information regarding sentiment polarities or emotions that may be used to help managers understand the main drivers of satisfaction or dissatisfaction discussed by consumers.
The current chapter explores the methodologies used to handle the large amounts of unstructured data such as those expressed in e-WOM comments. Although many techniques have been used to treat such data, the current chapter summarizes the techniques that have been used to successfully transform e-WOM into a set of structured patterns. Such patterns may then be used to help decision makers in devising strategies to better meet consumer needs.

ACKNOWLEDGMENT (Optional)

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Arora, M., & Kansal, V. (2019). Character level embedding with deep convolutional neural network for text normalization of unstructured data for Twitter sentiment analysis. *Social Network Analysis and Mining*, *9*(1), 12.

Arrami, S., Oueslati, W., & Akaichi, J. (2017). Detection of Opinion Leaders in Social Networks: A Survey. In *International Conference on Intelligent Interactive Multimedia Systems and Services* (pp. 362-370). Springer.

Atchison, J. & Shen, S. M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2), 261–272.

Bamakan, S. M. H., Nurgaliev, I., & Qu, Q. (2019). Opinion leader detection: A methodological review. Expert Systems with Applications, 115, 220-222.

Bharti, S. K., Pradhan, R., Babu, K. S., & Jena, S. K. (2017). Sarcasm analysis on twitter data using machine learning approaches. In *Trends in Social Network Analysis* (pp. 51-76). Springer.

Bilici, E., & Saygın, Y. (2017). Why do people (not) like me?: Mining opinion influencing factors from reviews. *Expert Systems with Applications*, *68*, 185-195.

Bilro, R. G., Loureiro, S. M. C., & Guerreiro, J. (2018). 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), 1-25.

Bing, L. (2019). Opinion Mining, Sentiment Analysis, and Opinion Spam Detection. Retrieved from https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, *1*(1), 17-35.

Blei, D. M. & Lafferty, J. D. (2009). Topic models. In Text mining: classification, clustering, and applications (pp. 71). Chapman & Hall, CRC Data Mining and Knowledge Discovery Series.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3, 993–1022.

Caro, L. Di, & Grella, M. (2013). Sentiment analysis via dependency parsing. *Computer Standards & Interfaces*, 35(5), 442–453.

Chen, Y. C., Hui, L., Wu, C. I., Liu, H. Y., & Chen, S. C. (2017, August). Opinion leaders discovery in dynamic social network. In *Ubi-media Computing and Workshops (Ubi-Media), 2017 10th International Conference on* (pp. 1-6). IEEE.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug), 2493-2537.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W. & Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407.

Delen, D., & Crossland, M. D. (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications*, 34(3), 1707–1720.

Farias, D. H., & Rosso, P. (2017). Irony, sarcasm, and sentiment analysis. In *Sentiment Analysis in Social Networks* (pp. 113-128). Morgan Kaufmann.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software*, 25(5), 1–54.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4), 82.

Fellbaum, Christiane (2005). *WordNet and wordnets*. In: Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics, Second Edition (pp. 665-670). Oxford: Elsevier.

Griffiths, T. & Steyvers, M. (2007). *Probabilistic topic models*. In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Ed.), Handbook of Latent Semantic Analysis. Hillsdale, NJ: Erlbaum.

Grün, B., & Hornik, K. (2011). topicmodels : An R Package for Fitting Topic Models. *Journal of Statistical Software*, 40(13), 1–30.

Hamilton, W. L., Clark, K., Leskovec, J., & Jurafsky, D. (2016). *Inducing domain-specific sentiment lexicons from unlabeled corpora*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing (pp. 595). NIH Public Access.

Hassler, M., & Fliedl, G. (2006). *Text preparation through extended tokenization*. Data Mining VII: Data, Text and Web Mining and their Business Applications, 37, 13–21.

Hearst, M. (1999). *Untangling text data mining*. Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (pp. 3–10).

Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet?. *Journal of interactive marketing*, *18*(1), 38-52.

Hofmann, T. (1999). *Probabilistic latent semantic indexing*. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 50–57).

Hristovski, D., Peterlin, B., Mitchell, J. A., & Humphrey, S. M. (2005). Using literature-based discovery to identify disease candidate genes. *International journal of medical informatics*, 74(2), 289–298.

Katyayan, P., & Joshi, N. (2019). Sarcasm Detection Approaches for English Language. In *Smart Techniques for a Smarter Planet* (pp. 167-183). Springer, Cham.

Kincl, T., Novák, M., & Přibil, J. (2019). Improving sentiment analysis performance on morphologically rich languages: Language and domain independent approach. *Computer Speech & Language*, *56*, 36-51.

Litvin, S. W., Goldsmith, R. E., & Pan, B. (2008). Electronic word-of-mouth in hospitality and tourism management. *Tourism management*, *29*(3), 458-468.

Liu, B., & Zhang, L. (2012). *A survey of opinion mining and sentiment analysis. Mining Text Data* (pp. 415–460). Springer US.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093–1113.

Minka, T. & Lafferty, J. (2002). *Expectation-propagation for the generative aspect model*. In Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence (pp. 352–359).

Mohammad, S. (2019). NRC Emotion Lexicon. Retrieved from http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.

Mukherjee, S., & Bala, P. K. (2017). Detecting sarcasm in customer tweets: an NLP based approach. Industrial Management & Data Systems, 117(6), 1109-1126.

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.

Nielsen, F.A. (2019). AFFIN Sentiment Lexicon. Retrieved from http://corpustext.com/reference/sentiment_afinn.html.

Porter, M. (1980). *An algorithm for suffix stripping*. Program: electronic library and information systems, 14(3), 130–137.

Sánchez, D., Martín-Bautista, M. J., Blanco, I., & Torre, C. J. D. La. (2008). *Text Knowledge Mining: An Alternative to Text Data Mining*. 2008 IEEE International Conference on Data Mining Workshops (pp. 664–672). IEEE.

Strahilevitz, M. & Myers, J. (1998). Donations to charity as purchase incentives: How well they work may depend on what you are trying to sell. *Journal of Consumer Research,* 24(4), 434–446.

Teh, Y. W., Newman, D., & Welling, M. (2007). A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. *In Advances in Neural Information Processing Systems* (pp. 1353-1360).

Vechtomova, O. (2017). Disambiguating context-dependent polarity of words: An information retrieval approach. *Information Processing & Management*, *53*(5), 1062-1079.

ADDITIONAL READING

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.

Guerreiro, J., & Moro, S. (2017). Are Yelp's tips helpful in building influential consumers? *Tourism Management Perspectives*, *24*, 151-154.

Santos, C. L., Rita, P., & Guerreiro, J. (2018). Improving international attractiveness of higher education institutions based on text mining and sentiment analysis. *International Journal of Educational Management*, *32*(3), 431-447.

KEY TERMS AND DEFINITIONS

e-WOM: All communication that is shared with peers through Internet-based technologies about the users opinion of goods, services, brands, or experiences.

Graph Mining: A set of techniques to extract and discover non-trivial, previously unknown and useful patterns from graph structures such as online social networks.

Natural Language Processing: A set of techniques based on many different disciplines such as computer science, artificial intelligence and linguistics, that allows computers to understand the human language.

Sentiment Analysis: The use of semi-automated techniques such as text mining, natural language processing and semantic rules to classify text according to its sentiment polarities or according to a sentiment scale.

Text Mining: The discovery of non-trivial, previously unknown and potentially useful information from text.

Topic Models: Topic models are a set of algorithms that uncover the semantic structure of a collection of documents based on a Bayesian analysis.