Abstract

Purpose

User-generated social media comments can be a useful source of information for understanding online corporate reputation. However, the manual classification of these comments is challenging due to their high volume and unstructured nature. The purpose of this paper is to develop a classification framework and machine learning model to overcome these limitations.

Design/methodology/approach

The authors create a multi-dimensional classification framework for the online corporate reputation that includes six main dimensions synthesized from prior literature: quality, reliability, responsibility, successfulness, pleasantness and innovativeness. To evaluate the classification framework’s performance on real data, the authors retrieve 19,991 social media comments about two Finnish banks and use a convolutional neural network (CNN) to classify automatically the comments based on manually annotated training data.

Findings

After parameter optimization, the neural network achieves an accuracy between 52.7 and 65.2 percent on real-world data, which is reasonable given the high number of classes. The findings also indicate that prior work has not captured all the facets of online corporate reputation.

Practical implications

For practical purposes, the authors provide a comprehensive classification framework for online corporate reputation, which companies and organizations operating in various domains can use. Moreover, the authors demonstrate that using a limited amount of training data can yield a satisfactory multiclass classifier when using CNN.

Originality/value

This is the first attempt at automatically classifying online corporate reputation using an online-specific classification framework.

Keywords

Banking industry Neural networks Social media Machine learning Online corporate reputation

Citation

Rantanen, A., Salminen, J., Ginter, F. and Jansen, B.J. (2019), "Classifying online corporate reputation with machine learning: a study in the banking domain", *Internet Research*, Vol. 30 No. 1, pp. 45-66. https://doi.org/10.1108/INTR-07-2018-0318

Download as .RIS

Publisher: Emerald Publishing Limited

Introduction

With the rapid development of information technology, consumer influence has grown and their role has shifted from passive receivers to active information producers (Heinonen, 2011; Owusu *et al.*, 2016). User-generated content (UGC) disseminates faster, cheaper and more widely than ever before, affecting the reputation of companies (Kim and Kang, 2018; Zhang *et al.*, 2011). Social media has rapidly become a valuable source of opinions (Belbachir and Boughanem, 2018) and recommendations on products and brands (Okazaki *et al.*, 2014; Xiao *et al.*, 2018), and eConsumers increasingly rely on online recommendations given by “who they know” rather than the content produced by the company (Steffes and Burgee, 2009). For this reason, it is vital for companies to monitor how their brands are discussed “in the wild” (Gensler *et al.*, 2015).

By analyzing UGC, companies can gain valuable information about why customers like or dislike a company’s products and services (Li and Li, 2013). Opinions published on social media can represent authentic information voiced directly by consumers (Presi *et al.*, 2014), as they tend to be produced without external pressure (Canhoto and Padmanabhan, 2015). Therefore, by studying UGC, different results can be obtained more readily than with traditional survey research. Additionally, customer information from social media can be collected in real-time and more cost-effectively than with more traditional data collection methods (Salminen, Şengün, Kwak, Jansen, An, Jung, Vieweg and Harrell, 2018). Social media platforms provide access to timely data, making it possible for companies to monitor their online corporate reputations (Gensler *et al.*, 2015).

For example, customers express their dissatisfaction on the internet to other users (Presi *et al.*, 2014), providing an opportunity for companies to stay alert to customer experiences and to manage their reputations (Salminen and Degbey, 2015). Naturally, there are also shortcomings in relying exclusively on social media comments, as there are a variety of issues that raise questions about their reliability and their variety for use within the domain of online corporate reputation. For example, there is the well-known issue of fake accounts (Shu *et al.*, 2017), bots (Davis *et al.*, 2016) and adversarial behaviors (Zaharna and Uysal, 2016). There is also the more nuanced aspect of people expressing themselves differently online than they do in face-to-face encounters (Lee and Pang, 2014), such as when people speak more hatefully online (Salminen, Almerekhi, Milenković, Jung, An, Kwak and Jansen, 2018). Even with these factors, social media data adds a degree of customer insight that was not possible before the widespread use of online social media.

It is important to note, though, that the increased volume of social media information makes it exceedingly difficult to monitor online corporate reputation. For example, a company may be discussed in thousands to millions of social media comments per day, week, or month. Manually analyzing vast amounts of online comments is a time-consuming and costly approach for monitoring the online reputation of a company. This research explores a means of automation by asking the question of how to automatically detect and classify online corporate reputation. Corporate reputation has traditionally been studied using questionnaires (Fombrun *et al.*, 2000), but novel automated methods based on machine learning offer great opportunities to process big data (Tirunillai and Tellis, 2014). Automatic classification of corporate reputation dimensions thus poses a contemporary challenge for companies and other organizations, even if they are not actively engaged in social media because the discussion is present whether companies react to it or not (Salminen and Degbey, 2015).

To address this challenge of user-generated social media comments, researchers started to extract information from online text content (Gensler *et al.*, 2015). Text mining methods have been used in analyzing brand sentiment (e.g. Mostafa, 2013) and brand image (e.g. Gensler *et al.*, 2015; Netzer *et al.*, 2012), but no prior study, to our knowledge, has used UGC and machine learning methods to investigate corporate reputation using an online-specific taxonomy. Earlier efforts utilizing machine learning for reputation and brand classification have focused mainly on classifying positive, neutral and negative dimensions (e.g. Cambria *et al.*, 2013; Mostafa, 2013), while ignoring more complex dimensions, such as quality and trust, that are seen essential for the brand reputation concept (Aaker, 1997).Therefore, to understand online corporate reputation, the applied classification frameworks need to be more comprehensive than a simple sentiment analysis involving nuanced aspects of language (Ruthven, 2019).

In this research, we develop a comprehensive classification framework for online corporate reputation by synthesizing prior works in the fields of corporate and brand reputation. After developing a multi-dimensional classification framework, we then apply machine learning, namely convolutional neural network (CNN) to detect and to classify the reputation dimensions from real user comments posted online. The intuition is that the CNN we develop can differentiate between signal and noise when detecting reputation dimensions. We validate our approach by classifying social media mentions for two Finnish banking companies.

In the following section, we review the related literature. After that, we develop and test the online corporate reputation framework and test it using a machine learning model. Finally, we discuss our experiences in applying machine learning for this task and the implications for future research and practical implementation.

Literature review

Defining online corporate reputation

Even though corporate reputation has been widely studied in the academic literature, there is no unequivocal definition of it. For example, Dowling (2016) lists 50 distinct reputation definitions between 1983 and 2014, out of which no definitive definition has emerged. In a similar vein, Chun (2005) notes that some definitions are overlapping and others are contradicting.

However, the various views of corporate reputation seem to contain a certain core element, namely that corporate reputation is a reflection of the company to insiders and especially outsiders (i.e. stakeholders) of the said company, and is linked closely to concepts of image and identity (Chun, 2005; Dowling, 2016). Therefore, we focus on this core aspect of corporate reputation, leaving investigating other nuanced aspects to future research.

Most researchers define corporate reputation as a collective concept (Abratt and Kleyn, 2012; Argenti and Druckenmiller, 2004; Chun, 2005). According to the collective definition, corporate reputation is based on a shared understanding of a company by its stakeholders, i.e., people associated with the company (Walsh and Beatty, 2007).For example, Fombrun *et al.* (2015) define a company’s reputation as the sum of stakeholders’ perceptions. Stakeholders evaluate the company’s products, communications, operations, financial situation (Chun, 2005) and interaction of customers, managers, employees, business representatives (Walsh and Beatty, 2007) with other people or groups linked to the company (Abratt and Kleyn, 2012).

Online corporate reputation can be seen as the extension of corporate reputation in the online environment (Dutot and Castellano, 2015). For example, Alwi and Da Silva (2007) and Christodoulides and De Chernatony (2004)maintain that online reputation can be evaluated using the same criteria as traditional reputation. In turn, Dutot and Castellano (2015) emphasize interactivity and trust, the credibility of data sources, as well as security and confidentiality of transactions as online-specific attributes. Online stakeholders base their assessments on a company's website (Dutot and Castellano, 2015) and social media content (Dutot *et al.*, 2016). Thus, online corporate reputation is not defined solely by what a company does or says, but by social media users voicing their opinions about brands (Floreddu *et al.*, 2014). These specific internet users can encompass current and prospective customers, along with those who may have heard of the company via eWord-of-Mouth (eWOM) (Jansen *et al.*, 2009). Additionally, these internet users may be stakeholders other than customers, such as current and former employees. This online aspect offers new challenges for companies in managing the conversation around corporate reputation. On the other hand, the internet offers companies a new way of listening to stakeholders and to tracking online discussions (Gensler *et al.*, 2015). By taking part in online conversations, a company can influence its reputation's development (Salminen and Degbey, 2015). Therefore, online corporate reputation is associated with increased loss of control and increased need for active monitoring.

Measurement of corporate reputation

Traditionally, a one-dimensional scale is used to measure reputation and to evaluate if the reputation is positive ("good") or negative ("bad") (Chun, 2005). This method is the simplest way to understand reputation, but it excludes the reasons why a company has a worse or better reputation than its rivals (Davies *et al.*, 2004). Different companies may have the same overall assessment of their reputation for "goodness," but their qualities may have different characteristics (Chun, 2005). Therefore, researchers believe reputation has a multi-dimensional nature (Dutot *et al.*, 2016; Walsh and Beatty, 2007).

Multi-dimensional reputation scales include Fortune magazine's annual America's Most Admired Corporations (AMAC) index, reputation quotient scale (RQ) by Fombrun *et al.* (2000) and Walsh and Beatty's (2007) customer-based reputation scale (CBR). The AMAC index focuses specifically on corporate management's point-of-view in assessing a company's reputation and customer-based scale in the customer perspective. RQ also considers other corporate stakeholders, such as employees and society. Based on RQ, the RepTrak scale was formed to give a better understanding of the elements of reputation that lead to emotional affection toward a company (Fombrun *et al.*, 2015).Reliability and validity of the RepTrak scale have been tested for various stakeholders' point-of-views such as customers, investors, the general public and key opinion leaders (Fombrun *et al.*, 2015; Lee *et al.*, 2018).

As shown in Table I, the instruments have many common dimensions, but there are some differences. In the AMAC survey, CEOs and analysts express their views on Fortune 500 and Fortune 1,000 companies' reputations (Chun, 2005).Respondents evaluate business reputation based on eight criteria: quality of management, quality of products and services, innovation, long-term investment value, financial stability, employee skills, use of company resources, and social and environmental responsibility (Davies *et al.*, 2004). In RQ, stakeholders evaluate the company according to its emotional appeal, products and services, vision and leadership, work environment, social and environmental responsibility, and financial performance (Fombrun *et al.*, 2000). The dimensions of

CBR include customer orientation, being a good employer, reliability, financial strength, product and service quality, and social and environmental responsibility (Walsh and Beatty, 2007). The criteria of RepTrak include products and services, innovation, workplace, management, citizenship, leadership and performance (Fombrun *et al.*, 2015). The AMAC index has three dimensions related to finances, which are combined into a single dimension in the other instruments. Financial performance is evaluated based on productivity, financial stability, competitiveness and future prospects (Fombrun *et al.*, 2000; Walsh and Beatty, 2007).

Among the stakeholders, investors are interested in a company's financial structure and performance (Gray and Balmer, 1998) as they evaluate a company based on investment potential (Dowling, 2016). Customers may find it harder to assess the company's financial performance (Walsh and Beatty, 2007). They are more interested in the quality and reliability of the company's products and services (Gray and Balmer, 1998). The AMAC and the RepTrak instruments have their own dimension for innovation, which has been integrated into two other instruments under the dimension of products and services (Fombrun *et al.*, 2000; Walsh and Beatty, 2007). The RQ and the CBR instruments evaluate the innovativeness, reliability, and quality of products and services. A company's products can be physical, and the services can take place at a physical office, or products and services can be in digital form. In both cases, the products and services may be evaluated online, affecting a company's online corporate reputation.

The management dimension can be found in the AMAC index, in the RQ, and the RepTrak instruments. The leadership dimension of the RepTrak instrument assesses the quality of a company's management and the clarity of the vision (Fombrun *et al.*, 2015). Additionally, the vision and management dimension of RQ evaluate the identification and utilization of possibilities (Fombrun *et al.*, 2000). The AMAC index measures only the quality of leadership. The employer dimension of the CBR estimates the quality of management, employees' skills and the way a company treats its employees (Walsh and Beatty, 2007). In RepTrak, caring for employees and equal treatment of employees are under the workplace dimension (Fombrun *et al.*, 2015). The work environment dimension of RQ measures the quality of management, the goodness of a company as an employer, and employees' skills (Fombrun *et al.*, 2000). Fortune's AMAC index has an additional employee competence dimension that measures a company's ability to attract, develop, and keep skilled employees. All four measurement instruments, therefore, measure the quality of management and employee skills in one way or another.

Customer orientation and emotional appeal have common features. Customer orientation implies that a corporation cares for its customers and their needs and treats them fairly (Walsh and Beatty, 2007). Emotional appeal measures how good a feeling about a company is formed and how a company is appreciated, respected or trusted (Fombrun *et al.*, 2000). One can think that if a company is customer oriented, it has a higher emotional appeal for its customers. The AMAC index or the RepTrak instrument do not have this type of dimension.

The AMAC measurement instrument, which focuses on a company's financial performance, has also received criticism (Chun, 2005; Davies *et al.*, 2004). Although financial indicators can explain changes in reputation, some of a company's reputation will inevitably remain unexplained if only such indicators are used (Chun, 2005). In addition to financial performance, emotional factors can determine how respondents perceive a company (Davies *et al.*, 2004). Also, Fortune's instrument has been criticized for only considering senior management's and economic analysts' views about corporate reputation without considering customer perspective (Walsh and Beatty, 2007). Nonetheless, it has often been used to measure a customer-centric reputation (Sarstedt *et al.*, 2013). The CBR instrument, however, only considers customers' point-of-view in assessing the reputation of a company; although, it may be difficult for a customer to estimate, for example, a company's goodwill as an employer, which is one of the instrument's dimensions. The emotional appeal dimension of RQ has been criticized because it more emphatically measures the results reputation has for a company rather than the property of reputation itself (Sarstedt *et al.*, 2013).

We consider the strengths and weaknesses of different reputation instruments when creating an online corporate reputation instrument. In the following, we synthesize the dimensions and items of the said instruments into a new instrument that we use for classifying online corporate reputation. The aims for this classification are to maintain the multidimensionality of the prior measurements and to consider several stakeholder perspectives.

Methodology

Research context

The two largest banks in Finland, Nordea and OP Financial Group, were selected for this study's context. The OP Group is the largest banking operator in Finland, consisting of about 180 independent cooperative banks with their central entity being owned by the OP Cooperative with its subsidiaries and related corporations. The group has 1.4m customer-owners, and its activities are banking, insurance and asset management. Nordea, in turn, is one of the ten largest banks in Europe. Its shares are traded on the OMX Nordic Exchange in Stockholm, Helsinki, and Copenhagen. Nordea has 10m private customers in the Nordic countries and 0.6m corporate and institutional customers.

Banks form an interesting subject for studying online corporate reputation as the banking industry has undergone a major digital revolution in recent times. New digital services are constantly developed, and efforts are made to serve customers online instead of via physical appointments. Different social media platforms have also become part of banking's customer relationship management (Mousavian and Ghasbeh, 2017). Both Nordea and OP Group have brand ambassador programs through which employees are encouraged to participate in social media. In addition to employees, Nordea's and OP bank's managers are also active on Twitter. For example, OP Group's former managing director Reijo Karhinen[1] is at the time of writing one of the most followed Finnish business executives on Twitter in Finland. Nordea is an international bank, but this research focuses only on Nordea's online reputation in Finland.

Nordic media has focused on the reputation of banks in recent years. In 2016, Nordea received considerable negative publicity due to news of tax evasion. In September 2016, several magazines published the results of a brand reputation survey conducted by T-Media in which Nordea was listed among the worst companies (Arola, 2016). Nordea's reputation fell in all areas, which included governance, finance, management, innovations, interaction, products and services, workplace and responsibility (Vänskä, 2017). The results of the annual banking and a financial survey (EPSI Rating Group, 2016) published in October 2016 also showed a decrease in Nordea's customer satisfaction, especially among private customers. The research material is part of 2016 social media conversations, so these up-to-date events and study outcomes published by the media may have impacted conversations about the banks at the time of data collection.

Data collection

Data collection for this research involves two social media platforms, Suomi24 (https://keskustelu.suomi24.fi/) and Twitter. The main difference between these platforms is that Suomi24 is an online discussion forum, and Twitter is a microblog. Twitter allows creating, commenting and sharing up to 280-character posts, i.e., tweets. The purpose of tweets is to share information, news, opinions and complaints, or to provide details of a user's daily life (Smith *et al.*, 2012). On Twitter, consumers also share information about their interactions with companies, making information about online corporate reputation readily available. Hashtags, marked with "#,"are used on Twitter to annotate the topic of the tweets. One purpose of hashtags is to make tweets easier to find with the Twitter search feature (Pönkä, 2014). In this study, hashtags and usernames that mention the name of the banks being investigated are used for collecting Twitter data.

Suomi24 is the largest discussion forum in Finland. It includes some 3,000 sub-topics from everyday problems to shopping experiences and entertainment. It is possible to write anonymously with a pseudonym or with a registered username. Due to anonymity, forum discussions can provide different information than other channels. For example, users may express opinions that would not be disclosed face-to-face. The Suomi24 forum has 832,000 visitors per week. The gender of users is evenly distributed: 51 percent for men and 49 percent for women. The age of visitors is between 15 and 49 years (Lagus *et al.*, 2016).

The Suomi24 data set contains discussion forum posts between 2001 and 2014 and was obtained via the Korp Language Bank of Finland (www.kielipankki.fi/language-bank/), which is a linguistic data sets repository. The Suomi24 discussion forum messages have been made available for researchers by the company that owns the discussion forum. The keywords used to identify the material were "OP Financial Group" and "Nordea." The written data collection script for Suomi24 considers different forms of brand names and both upper and lower cases. The publicly available Tweepy library (www.tweepy.org) was used to collect Twitter data from its application programming interface. Table II lists the keywords used for collecting Twitter data.

Tweets were collected from both banks containing either the customer service's account or the main account's Twitter username. The purpose was to collect tweets addressed to these usernames by users who have written a comment addressed to the bank. Customer service accounts are designed for customer advice. For customer service accounts, users usually address questions or feedback to the bank, but main bank accounts may be used in the same manner. The collection was set up for both usernames, as user comments about banks may be tweeted using one or the other account. In addition to usernames, hashtags, i.e., identifiers used by people about these banks, were manually searched and used to collect the data.

Tweepy collects all the material, regardless of whether it is an original tweet or a retweet. In the data, retweets have an acronym RT in front of the tweet. All tweets with this abbreviation were deleted because they cannot be classified as content produced by users. Additionally, maintaining retweets in the material would cause the same comment to be duplicated. Table III summarizes the data collected from social media that was used in the study.

Numbers in the Twitter column do not include retweets, and the duplicates in Suomi24 data were also deducted. The conversation about banks was active on Twitter, so data were collected to a considerable extent. In the discussion forum Suomi24, Nordea was more exposed. Interestingly, the number of discussions about Nordea in relation to the OP Financial Group increased in Suomi24 in 2016, since during the year, data about Nordea have accumulated more than in 2001–2014 combined. This increase may be the cause of Nordea's negative media exposure due to the Panama scandal at that time.

Development of classification framework for online corporate reputation

When analyzing multi-dimensional reputational scales, a clear convergence between the dimensions of different dimensions was observed. These include quality, innovativeness, social responsibility and successfulness, which implies good financial results. In addition to these, “pleasantness” was the dimension to be identified, which corresponds to the “customer orientation” and the “emotional appeal” of multi-dimensional scales. “Reliable” and its close attributes were dispersed in different dimensions on the reputation scales. Since the attributes of reliability did not have a clear placement in other dimensions, a new dimension was created for it. The dimensions of the company’s online corporate reputation are “quality,” “pleasantness,” “innovativeness,” “reliability,” “responsibility” and “successfulness.”

We found that the construct of reliability that was not explicitly mentioned in prior works surprising, as it is a construct used in a variety of corporate measures, including advertising (Boateng and Okoe, 2015), trust (Van Der Merwe and Puth, 2014), marketing (Eteokleous *et al.*, 2016), ethics (Markovic *et al.*, 2018) and other aspects of the corporation; therefore, we explicitly included it in our instrument.

In the multi-dimensional reputation scales, the RQ, the CBR and the RepTrak scales’ dimensions are divided into questionnaire items that specify what dimensions are measured. The AMAC scale differs from other scales so that the dimensions of the scale are not public (Sarstedt *et al.*, 2013). In Table IV, the attributes of the RQ, the CBR and the RepTrak are classified under the online corporate reputation dimensions.

For classification, it is crucial to acknowledge that online comments can have a positive or negative sentiment (Mostafa, 2013). Therefore, we split the dimensions into (+) and (–) classes. For example, a comment “this company can clearly come up with creative solutions” would be classified as Innovativeness(+), whereas “this company rarely brings anything new to the market” would be Innovative(−). Therefore, the final number is 6 × 2=12 reputational classes.

In addition to the reputational classes, a Neutral class (Feldman, 2013) was used to classify neutral data that do not express any opinion on the company, and that do not fit into any reputational class. This was done because analyzing a sample of the collected comments revealed that only about 13 percent of the comments contained a user’s opinion about the bank. The rest of the data were neutral or noisy comments that do not represent any opinion. Noise is a common condition when dealing with UGC (Netzer *et al.*, 2012), so it is important to consider its existence when creating a classification for machine learning.

Creating the training data

One researcher and two research assistants independently and manually coded the data. Both research assistants were graduate students majoring in marketing in the Turku School of Economics. They were familiar to one of the researchers, but neither was familiar with the research, research methodology or the field of studying reputation. One of the research assistants has several years of work experience in the banking sector, but that experience was not in either of the two banks studied in the research. Assistants could thus be considered objective in terms of the research.

The research assistants were given a spreadsheet with classification rules, including keywords and example comments for each class. In addition to the 12 classes of reputation, the spreadsheet contained the Neutral class where assistants were instructed to place all the comments they found not to belong to any reputational class. The assistants were also given a document containing written instructions. The instructions provided a brief description of what the assistant should do and what company reputation means in this context. It was highlighted that a company’s reputation is not influenced only by a company itself but also by product reviews, services, employees, managers, business representatives, customers, communication, business activities, financial situation and other factors related to business and people. Moreover, a list of indicative words for each class was created to facilitate the coding (Table V).

The online context provides some challenges we considered when manually annotating the training data. For example, in one context, the sentence can be interpreted as positive and as negative in the other (Ortigosa *et al.*, 2014).Moreover, one part of the text can express more than one sentiment and refer to more than one object, which creates uncertainty for sentiment classification. Finally, critical reviews of brands often also have irony and sarcasm that is difficult to identify, which makes it challenging to classify sentiments (Canhoto and Padmanabhan, 2015). We considered the first issue by instructing the annotators to consider the sentiment of the comment carefully. For the second issue, we asked them to use as many classes as needed. For example, comments with two opinions:

> Manager of OP Financial Group is the absolutely the best and friendliest customer service person.

This comment would fit both Quality(+) and Pleasantness(+) classes, as the word “best” represents by the vocabulary, quality and the word “friendly” agreeableness. In such situations, the classification decision could be based on the overall picture that is generated from the comment (Canhoto and Padmanabhan, 2015). In the case of this comment, both classes are equally true, so it was decided to place the comment on both classes. In manual classification, this is possible, but in multiclass classification, the machine can only make the classification into one class[2]. Several comments classified into their classes, but the classification of some of the comments produced challenges:

> OP Financial Group is ok, but Nordea sucks ass.

In the case of this comment, the classification depends on which bank’s point-of-view classification is done. In the case of OP Financial Group, this can be classified as Quality(+). The word “ok” does not really sound good quality, but compared to Nordea, OP Financial Group is better in quality than Nordea. In the case of Nordea, this comment is classified as Quality(−).

Critical reviews of brands often have irony and sarcasm, which can be difficult to identify (Canhoto and Padmanabhan, 2015). Some sarcastic comments could be detected by manual classification:

> OP Financial Group is an honest, Finnish bank, not even the laws apply to it, it can do what it likes, and the charges will not be raised against it.

> According to Nordea, cash machines are not museum objects, but modern equipment.

Based on the first sentence in the first comment, the comment could be classified as the Reliability(+) class, because OP Financial Group is described as an honest bank; however, the rest of the sentence makes it clear that OP Financial Group does not comply with the laws. Consequently, a sentence can be interpreted as meaning that OP Financial Group operates criminally, and the comment is classified as Responsibility(−). The latter comment can be interpreted as meaning that Nordea is old-fashioned because, in Nordea, cash machines are considered modern equipment instead of museum artifacts. Thus, the comment is classified as an Innovativeness(−) class.

After the researcher completed classifying the commentary data, the two assistants classified the material. It took the assistants approximately two hours to independently make the classifications in Excel. The assistants were provided an Excel file with the classification rules and words, which provided guidelines for making the classification. The classification was performed in a spreadsheet. In the spreadsheet, the lines contained comments, and the columns contained reputational classes. The assistant’s task was to put the number one under class where they thought the comment belonged. If they felt that the comment belonged to more than one class, they had to choose the best class that they thought the comment would fit and to comment after the line where they would think it could fit also.

The assistants reported the spreadsheet was easy to fill. This facilitated the work of the researchers, as the spreadsheet was in an appropriate format for the calculation of inter-rater agreement. The agreement of the classifiers was measured with Fleiss’ Kappa because there were three raters (Bernard *et al.*, 2016). The accuracy of the test classification was 0.49, which implies average agreement among the classifiers. Invariably, a classification based on user’s judgments involves a great degree of judgment. Also, the risk of misinterpretation always exists when analyzing texts written by another person. Finally, having many classes makes classification tasks more challenging for human annotators. We coped with these facts by reviewing the comments in the training data one by one.

Thus, one of the researchers reviewed the labeled comments one by one. If both assistants, or even one of the assistants, agreed with the researcher about the comment’s class, the comment could be left to that class. If both assistants classified the comment differently from the researcher, the researcher considered whether the comment should be moved to a class proposed by assistants or retained in both classes. If all three had classified the comment into different classes, one of the researchers decided whether it should be moved to the Neutral class. There was no need to delete any comments because comments that were not classified in any of the reputation class belonged to the Neutral class. Table VI displays the training data after these efforts.

Developing and validating the machine learning model

In general, machine learning can take a supervised or an unsupervised form. Unsupervised learning is used when categories to which the machine should assign data are not known. In that case, the machine learning model analyzes the elements linking the data points and independently forms categories based on them (Bell, 2014). In this research, dimensions of online corporate reputation were first formed from the theory, after which training data were manually annotated and fed to the model. Therefore, we apply supervised machine learning.

Supervised machine learning is based on the idea that a system can be taught by showing it correct examples of training data from each class (Bell, 2014). Training data contains features that the machine learning model uses in text classification (Habernal *et al.*, 2014). For our text classification problem, manual coding is a necessary step for generating training data for the supervised machine learning model. After performing manually coding the data using the classification framework developed, the resulting training data is fed into the CNN and, after training, the whole data set is submitted to the neural network to classify. The classification’s success is verified from a random sample of comments.

The data were tokenized and lemmatized with the latest version of the Finnish dependency parser pipeline (Haverinen *et al.*, 2014), also available on the GitHub software repository (https://turkunlp.github.io/Finnish-dep-parser-neural). This parser is based on deep learning and achieves state-of-the-art performance on the Finnish language. Most importantly, for this study, the parser produces base forms of words (lemmas), reducing the data sparsity problem of the highly inflective Finnish language. To train the network, the comments are transformed into a numerical vector format that the network can utilize to identify various associations between class labels and training data.

The model builds on the CNN architecture, which is shown to be highly successful in recent work on text classification in other contexts (Lai *et al.*, 2015). The method's strength is based on inferring latent associations from the text to assign the most probable class among pre-defined classes (Joulin *et al.*, 2016). The CNN architecture is one of the most popular models for learning a fixed-length vector representation of the variable-length input document, which can be used in a subsequent classification layer. The CNN architecture, originally introduced in image recognition, is based on sliding a set of filters across all positions on the input document (see Figure 1). At each position in the document, each filter aggregates several subsequent words at the given position into a single activation value. The aggregation is implemented using the convolution operation between the input and the learned filter weights. The higher the value, the better "match" of the filter at the given position. Subsequently, the vector of maximal activations of each filter across the document constitutes the representation of the document for further classification. The length of this representation corresponds to the number of applied filters and is a hyper-parameter of the representation. Intuitively, each filter recognizes a pattern of words and reacts with a high activation when the pattern occurs in the document. In this sense, the filters act as learned feature extractors. Note that here "pattern" and "feature" are not to be interpreted in a strict, traditional present/absent sense, as all representations are continuous, and the matching is therefore fuzzy. Alternative representations would include, e.g., recurrent networks such as the LSTM model, but in our preliminary evaluations, these under-performed compared to CNN, likely due to the limited amount of training data available. Therefore, we concentrated on the CNN architecture.

Furthermore, the model employs pre-trained word and lemma vector space embeddings induced from the entire corpus of the Suomi24 Finnish online discussion forum using the popular word2vec method (Mikolov *et al.*, 2013). In this manner, the model utilizes raw, unannotated textual data to partially offset the limited training data size. The techniques, including tokenization, lemmatization and vectorization are commonly-used natural language processing methods (Gil *et al.*, 2013).

First, the internet forum and Twitter posts are turned into sequences of vectors corresponding to the pre-trained embeddings of words and lemmas – each word or lemma being represented using a single vector. Subsequently, these sequences are passed through CNN layers with window widths 1 and 2, and are aggregated using the standard max-pooling operator. In non-technical terms, these layers turn the sequence of word and lemma vectors into a single vector composed of maximally activated 1–2-word long filters learned by the network. This single vector represents the post for further classification, and during training, the neural network learns to build a representation suitable for the task.

Finally, this single vector representation is provided to three separate classification layers: the first that only predicts the raw sentiment (positive, negative, neutral); the second, which predicts the dimension without the sentiment; and the third, which predicts the combination of sentiment and dimension. These output layers are trained simultaneously, all jointly affecting the underlying convolutional layers. Thus, all dimensions with a positive sentiment can benefit from each other through the pure-sentiment output layer; yet, we do not need to predict the dimension and sentiment isolated from each other, making unjustified independence assumptions. This type of multi-task training is one of the notable advantages of deep learning architectures. Since the distribution of the classes is highly imbalanced, a class weighting scheme inversely proportional to the number of training examples is used, placing larger weight on the smaller classes. This prevents the classifier from ignoring the rare classes.

The model has numerous hyper-parameters, which, to a varying degree, affects its classification accuracy. Some parameters, for instance, the number of convolutional filters and layer widths, were set during the model's initial development. To optimize the most critical parameters, learning rate, drop-out rate and the L2 weight regularization parameter, we apply the RBFopt method (Costa and Nannicini, 2014). The learning rate controls the magnitude of updates to the neural network weights, and an inappropriate value tends to result in a network with grossly sub-optimal performance. The drop-out and L2 parameters, in turn, control the model's regularization and prevent overfitting during training. The RBFopt method implements a search through the parameter space, identifying the parameter settings leading to the best performance on the development set. The advantage of RBFopt compared to a brute-force grid search is its ability to more rapidly identify the "sweet spot" in the parameter space; it is therefore especially suitable to the relatively expensive neural network training. The optimal hyper-parameter values are set once and then used when training the ten final models. These models are evaluated based on their respective test sets not used to select the parameter values, thus avoiding an overly optimistic bias in the results. For replicability, the exact source code of the classifier as used in this study, the exact values of all parameters, as well as the data are available on GitHub (https://github.com/fginter/tw_sent_v2).

Findings

The accuracy of the classification is verified using a procedure whereby all available annotated data is randomly divided into training, parameter optimization and test set in 80:10:10 proportions. The available data are stratified, i.e., following the class distribution with a guarantee that every class represented is also in the test data. The training set is used to train the neural network, the development set is used to stop the training at the peak accuracy before the performance would start degrading due to overfitting, and the test set is then used to measure the final performance of the model. The success of the classification is verified from a random sample of comments from Nordea's and OP bank's Twitter and Suomi24 data. Since the amount of data is quite limited, we follow the standard technique of repeating the data split procedure ten times and averaging the results, reducing noise in the reported scores.

Testing the model, we find that 65.2 percent of Nordea's tweets and 61.7 percent of OP bank's tweets are correctly classified. In the case of Suomi24 data, 52.7 percent of Nordea's and 61.5 percent of OP bank's data are correctly classified. Out of the four data sets, Nordea's Suomi24 data were the weakest to classify. To investigate the model performance in detail, we compute a confusion matrix that shows the accuracy of each class prediction (Fawcett, 2006). Based on the confusion matrix, we can calculate the precision of the machine learning, recall, and F1 parameters. Precision measures how many positive class predictions are positive and recall measures how many real positives the machine predicted correctly. If both values are 1.00, the classifier is perfect; if close to zero, the classifier is highly inaccurate (Aghdam and Heravi, 2017). The F1 parameter represents the harmonic mean between precision and recall. The computed metrics are shown in Table VII.

As seen from Table VII, the strongest classes are Neutral, Innovativeness(+), Reliability(−), and Successfulness(+). Precision is strong in the class Reliability(−), but the recall is weaker, i.e., there was actually an even larger number of comments that should have hit this class, but they were incorrectly classified into other classes. For Responsibility(−), the situation was the opposite. The machine predicted the class to have more comments than it actually did, so the precision of the class was weaker than the recall. The weakest classes are Reliability(+), Successfulness(−), Responsibility(+), and Innovativeness(−). These classes contained only tens of comments, most of which were classified into the Neutral class. The training data, therefore, were not enough to classify these classes correctly. Looking at the confusion matrix (Table VIII), many observations would have belonged to pleasantness, but that was classified to the Neutral class. Moreover, Quality comments seem to have been mixed with Pleasantness comments, which is not surprising because there was some difficulty in distinguishing the two even in the manual coding. A look at the data reveals that most of the Quality(−) comments that were classified as Neutral deal with plans to change the bank or to avoid Nordea, probably due to the Panama Papers stir.

Discussion

In this research, we demonstrated an approach for automatically classifying several thousands of social media comments with a nuanced classification capturing various dimensions of corporate reputation. Reputation is of foremost importance for the development and maintenance of long-term relationships with stakeholders. Sudden reactions and comments online can strengthen or degrade a business’s reputation faster than ever before, highlighting the importance of automatic reputation monitoring and classification consumer-facing companies in all sectors.

From a theoretical point-of-view, we found that the reputational dimensions from earlier research do not entirely match with the topics and the language used by consumers on social media. This points to a possible weakness in prior work that used small sample sizes or samples that inadequately represented the online consumer and stakeholder populations. Our results suggest the concept of corporate reputation and the measurement instruments employed in prior work have not adequately reflected online corporate reputation. As our instrument was based on this prior work, our instrument also did not capture the concepts and dimensions appearing in the social media comments generated by the users.

Developing the training data by using the theoretically derived dimensions was challenging, as the dimensions are of a high level of abstraction and do not always represent the topics that online users are discussing concretely. This proposition is supported by the observed low inter-rater agreement score (Fleiss’ $\kappa = 0.49$). It seems that theoretically complex reputation classification schemes present challenges for obtaining an agreement between human coders. Because the machine learns from the training data, this indicates that simpler classification schemes are likely to perform better for both human and machine classification. A possible reason for this is the lack of abstract thinking or for the ambiguous nature of abstract notions, such as reliability and quality that may be difficult to adequately express in short social media comments.

Future work should, therefore, develop a corporate reputation classification scheme that is grounded in the actual online discussions. Using the actual comments rather than theory as a starting point, the online corporate reputation classification more accurately reflects the consumers’ use of language, concerns pointed out and overall topics of the commenting. Possibly, these concerns can be a company – or can be industry specific, but it is not clear whether a general, theory-based classification of online corporate reputation classification is more suitable than other methods, such as machine learning with open coding (cf. Salminen, Almerekhi, Milenković, Jung, An, Kwak and Jansen, 2018).Relatedly, a major future work effort would be to statistically validate the constructs of online corporate reputation via a series of pilot and then survey implementations. This could, perhaps, be done via crowdsourcing means with a focus on, as in the research presented here, the core aspect of online corporate reputation rather than on various nuanced definitions.

In our research, more consistent results could potentially have been achieved with fewer classes so that our assistants would have been able to follow the classification rules more closely. Future research should, therefore, aim at: reducing the number of classes for a simpler representation of corporate; ensure a match between classification and how users speak about brands; collect enough training data per class; experiment with different computational approaches, such as multi-label classification that enables us to capture several dimensions per comment; and conducting construct validity and reliability research for a more robust instrument to measure online corporate reputation.

It is apparent from our findings that corporate reputation is actually manifested in a myriad of ways, in a user-dependent manner, and can essentially be defined as what stakeholders and customers express about a corporation in a particular context at a given time. Therefore, online corporate reputation is a dynamic concept that is fluid over time, location, customers and stakeholders. As such, it may be difficult to develop a concise, tidy and robust theoretical definition, or a universally applicable classification framework. However, the approach of using machine learning for automatically classifying online corporate reputation shows clear promise and advantages.

Regarding practical implications, we encourage organizations to develop empirically “data-driven” classifications rather than relying on “theory-driven” conceptualizations of corporate reputation. By adopting the online corporate framework in their organizations for their own stakeholders with dimensions that are most relevant to their context, companies can achieve a substantial increase in speed and comprehensiveness over traditional reputation measurement methods. Overall, our findings encourage companies to create reputation classification frameworks that are: simple; unambiguous; and concrete to fully leverage machine learning. Automated bottom-up approaches enabled organizations to inductively generate more contextualized, detailed, and relevant pictures of their corporate reputations, empirically, and without the constraints of ill-fitting theoretical models.

To this end, we provide the following practical recommendations for machine learning projects for classifying online corporate reputation:

- Focus on the quality of the training data: to ensure the quality of the training data, which is crucial for the machine to pick up signals, focus on: creating a clear classification framework that captures the key dimensions of the classified phenomenon and contains minimal overlap between categories; providing clear instructions for manual annotators; and using several annotators and an iterative process to solve misunderstandings. In the process of creating training data, it is advised to use corporate reputation experts. Overall, technical parameter optimization cannot overcome a situation where the training data is flawed, whereas, with high-quality training data, even a simple classification algorithm can perform very well.
- Avoid theoretically abstract classes: one of the downsides of our classification framework was finding a match between the theoretically sound ideas about what corporate reputation, according to researchers, should consist of and the way people speak about companies online. In many cases, there may be a gap between the higher level of abstraction and the self-expression of online users. One remedy for this is to use methods such as open coding (Strauss and Corbin, 1998) to generate the classification framework from the data instead of relying on theories.

The classification framework, as well as the process for data collection, annotation of training data and classifier development, can be generalized to any domain and applied by any company or organization willing to invest resources in machine learning. For this reason, the approach presented here has real practical value for companies and organizations interested in automatically measuring their online reputation, but that find the manual investigation too time-consuming and costly.

As noted by several scholars, the market impact of automation and data-driven methods is enormous and transformative (Erevelles *et al.*, 2016; Loebbecke and Picot, 2015; Zerbino *et al.*, 2018). While this impact is discussed in other works that explain how big data, machine learning and artificial intelligence are shaping industries (Wang *et al.*, 2018), this research provided an example of the power of those technologies in one specific problem – one that many brands and organizations struggle with. The actual value of the approach presented here depends on whether companies would adopt it in real usage. Because there is widespread interest in artificial intelligence technologies among corporate decision makers (Deloitte, 2018), studies such as ours, that demonstrate the applicability of machine learning in real business problems, are called for.

Conclusion

This research is among the earliest to use machine learning for automatically classifying online corporate reputation. We tightly focus on solving a real problem that many, if not most, brands and companies face – the monitoring of what is said about them online. We demonstrate the applicability of computational techniques in automatically processing substantial amounts of social media mentions using a theory-based classification of a company’s online reputation. We also show that the automatic classification of online reputation is not a trivial problem, as reputational dimensions are subjective, and the concept itself involves many dimensions with a high degree of abstraction.