Communications of the Association for Information Systems

Volume 44 Article 43

6-2019

Off-The-Shelf Artificial Intelligence Technologies for Sentiment and Emotion Analysis: A Tutorial on Using IBM Natural Language Processing

Arthur Carvalho
Miami University - Oxford, arthur.carvalho@miamioh.edu

Adam Levitt
Miami University

Seth Levitt
Miami University

Edward Khaddam
University of Tennessee

John Benamati
Miami University

Recommended Citation

Carvalho, A., Levitt, A., Levitt, S., Khaddam, E., & Benamati, J. (2019). Off-The-Shelf Artificial Intelligence Technologies for Sentiment and Emotion Analysis: A Tutorial on Using IBM Natural Language Processing. Communications of the Association for Information Systems, 44, pp-pp. https://doi.org/10.17705/1CAIS.04443

Communications of the Association for Information Systems

Tutorial DOI: 10.17705/1CAIS.04443 ISSN: 1529-3181

Off-The-Shelf Artificial Intelligence Technologies for Sentiment and Emotion Analysis: A Tutorial on Using IBM Natural Language Processing

Arthur Carvalho
Farmer School of Business
Miami University
arthur.carvalho@miamioh.edu

Adam Levitt
Farmer School of Business
Miami University

Seth Levitt
Farmer School of Business
Miami University

Edward Khaddam
Haslam College of Business
University of Tennessee

John Benamati
Farmer School of Business
Miami University

Abstract:

Artificial intelligence (AI) rests on the premise that machines can behave in a human-like way and potentially solve complex analytics problems. In recent years, we have seen several off-the-shelf AI technologies that claim to be ready to use. In this paper, we illustrate how one can use one such technology, called IBM Natural Language Understanding (NLU), to solve a data-analytics problem. First, we provide a detailed step-by-step tutorial on how to use NLU. Next, we introduce our case study in which we investigated the implications of Starbucks' pledge to hire refugees. In this context, we used NLU to assign sentiment and emotion scores to social-media posts related to Starbucks made before and after the pledge. We found that consumers' sentiment towards Starbucks became more positive after the pledge whereas investors' sentiment became more negative. Interestingly, we found no significant relationship between consumers' and investors' sentiments. With help from NLU, we also found that consumers' sentiments lacked consensus in that their social media posts contained a great deal of mixed emotions. As part of our case study, we found that NLU correctly classified the polarity of sentiments 72.64 percent of the time, an accuracy value much higher than the 49.77 percent that the traditional bag-of-words approach achieved. Besides illustrating how practitioners/researchers can use off-the-shelf AI technologies in practice, we believe the results from our case study provide value to organizations interested in implementing corporate social responsibility policies.

Keywords: Artificial Intelligence, Sentiment Analysis, Corporate Social Responsibility, CRISP-DM.

This manuscript underwent peer review. It was received 06/19/2018 and was with the authors for 4 months for 2 revisions. Ming-Hui Huang served as Associate Editor.

1 Introduction

In recent years, we have seen the emergence of several technologies that focus on helping decision makers harness “big data”. Many such technologies use artificial intelligence (AI) techniques such as machine learning, natural language processing, and knowledge representation. Several IT companies, such as Google, IBM, Amazon, and Microsoft, now offer off-the-shelf AI technologies via their cloud platforms. These technologies greatly help one use powerful techniques to solve a variety of problems, such as recognizing objects in images, creating bots to interact with humans, and extracting sentiments from texts in an automated manner. We focus on this last topic in this paper.

Sentiment analysis belongs to the natural language processing (NLP) area, which, in turn, many researchers see as a subfield of AI (Russell & Norvig, 2010). Some other NLP tasks include text classification, machine translation, and text summarization (Chowdhury, 2003). Traditionally, appropriately using NLP techniques required considerable expertise in (computational) linguistics. Cloud-based, off-the-shelf technologies promise to make the usage process considerably easier. For example, users no longer need to perform extensive data preprocessing and modeling. Instead, one can often simply submit raw text to the cloud-based technology via an application programming interface (API), and the technology will do all the complex work and return an outcome (e.g., a sentiment score) to the user.

In this paper, we provide a detailed step-by-step tutorial on how to use one of the most prominent off-the-shelf AI technologies to perform sentiment analysis: natural language understanding. We believe this tutorial will have great value to practitioners and information systems (IS) researchers who wish to analyze textual data from different sources (e.g., social media) in an efficient and effective way.

In addition to providing a step-by-step tutorial on how to use IBM Natural Language Understanding (NLU), we also show how one can use this technology in practice when solving a real-life research problem. In our case study, we investigate the implications of a Starbucks’ corporate social responsibility policy; namely, its pledge to hire thousands of refugees worldwide. In this context, we use NLU to assign sentiment and emotion scores to Facebook posts about Starbucks before and after the pledge. By doing so, we can understand the changes in sentiments and emotions of potential Starbucks consumers towards the company. We believe the results from our case study provide value to organizations who are considering implementing corporate social responsibility policies.

2 Natural Language Understanding

One can use artificial intelligence (AI) to, among other things that its functions allow, create computerized models that simulate human thought processes to solve computational problems (Russell & Norvig, 2010). To do so, AI uses data mining, pattern recognition, signal processing, knowledge representation, natural language processing, and other well-established techniques. Although researchers have not come to a consensus about how to define AI or its scope, some common AI tasks include speech recognition, face recognition, and sentiment/emotion analysis. We focus on the latter task in this paper to illustrate how to use an off-the-shelf AI technology to perform sentiment and emotion analysis.

To assign sentiment and emotion scores to texts, we use IBM's technology called Natural Language Understanding (NLU). IBM hosts NLU, which forms part of the AI/Watson family of services, on its cloud platform. Note that, even though NLU belongs to the Watson family of services, it differs from the original Watson (a Q&A system) that won the game show Jeopardy! (Ferrucci et al., 2010; Ferrucci, Levas, Bagchi, Gondek, & Mueller, 2013). Instead, the NLU service evolved from the very popular AlchemyAPI after IBM purchased it (IBM, 2015).

Currently, one can use NLU in two distinct ways: 1) via a Web interface and 2) via an API. First, one can understand the results that NLU produces via the Web interface by trying the online demo (NLU, 2018). To illustrate NLU, consider the following Facebook comment about Starbucks’ refugee hiring policy that belongs to the data set from the case study we discuss in later sections:

> *To the Starbucks Corporation, as the wife of a uniformed service member, I want to applaud your pledge to employ refugees. Thank you for the ray of hope in a frightening time, and for demonstrating humanity in the realm of business.*

NLU can calculate sentiment and emotion scores and detect keywords, entities, categories, concepts, and semantic roles. Figure 1 displays some results from applying NLU to the above comment. Regarding sentiment scores, the figure shows that NLU assigns a very positive sentiment to the underlying text. Regarding emotions, NLU heavily features the feeling of joy present in the text. NLU can also identify the most important keywords in the comment, such as “uniformed service member”, “pledge”, and “refugees”. It can also retrieve key entities such as “Starbucks Corporation”. One can download all these results from the Web interface as JSON files.

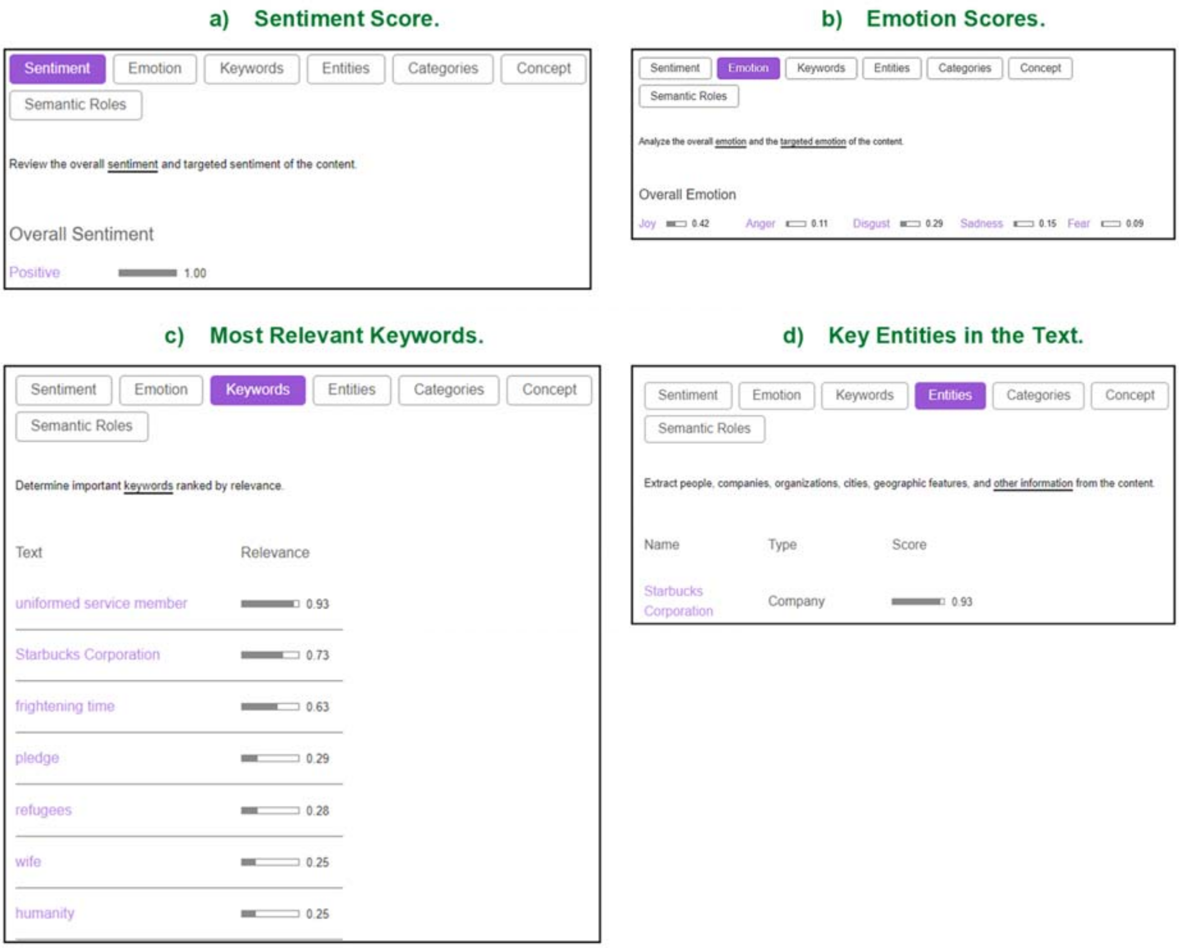


Figure 1. Illustration of an Application of NLU

One can use NLU’s Web interface in a project by manually pasting each text that one wants to analyze (one at a time) into NLU’s website and then downloading the produced outcomes as a JSON file. Clearly, this approach lacks practicality when one has to analyze many texts.

Second, one can use NLU via its API. APIs allow cloud platforms to provide services/data in a controlled manner (e.g., the platform knows how many queries a certain authenticated user has made, which allows the platform to set up the billing accordingly). Users traditionally request Web services using protocols such as SOAP and REST. The services often respond in file formats such as XML and JSON. NLU, in particular, uses the very popular REST/JSON combination. By using the API, one can create scripts that completely automate the process of submitting texts to NLU and deriving the desired responses (e.g., sentiment scores) from the returned JSON files. In Sections 2.1 and 2.2, we illustrate how one performs this process. Specifically, in Section 2.1, we describe how to obtain the required credentials to use NLU’s API. In Section 2.2, we demonstrate how to use the API.

2.1 Obtaining the Required Credentials

NLU belongs to the Watson family of services (recently rebranded as "AI"), which IBM's cloud platform hosts. As such, one needs to create an account on that platform first in order to use NLU. At the time of writing, one can do so by clicking on the button labeled "create a free account" at https://console.bluemix.net/. After creating an account and logging into the system, one can then click on the button labeled "catalog" to see the several cloud-based products that IBM offers, which range from computational infrastructure (e.g., database servers) to AI technologies. When selecting "AI", one will then be able to see the many AI technologies that IBM offers: from language-translation services to object recognition in images. Figure 2 illustrates the AI services that IBM currently offers. In this paper, we focus on the "Natural Language Understanding" service. After selecting that service, one can choose the desired pricing plan. At the time of writing, IBM offers three plans: 1) lite: the user can perform up to 30,000 queries a month for free using one pre-trained model, 2) standard: the user pays for the number of trained models (US$800 per model and at most US$0.003 per subsequent query), and 3) premium: the user can contact IBM directly to obtain compute-level isolation on the existing shared cloud platform and end-to-end data encryption.

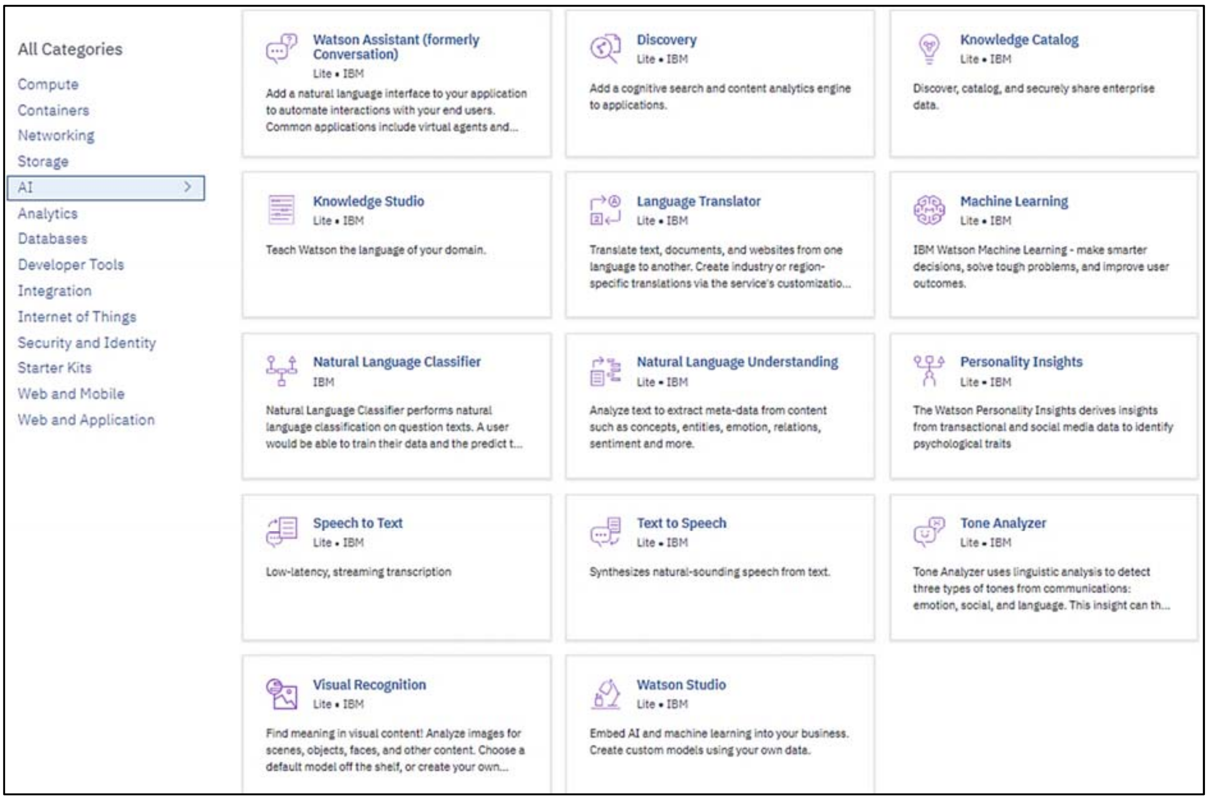


Figure 2. Some of the AI Services Currently Offered by IBM on Its Cloud Platform

We focus on the lite plan. After selecting the appropriate plan, the user can now see the created service "Natural Language Understanding" after clicking on the "dashboard" menu item. When selecting that service, the user will then be able to obtain the required credentials to use NLU's API. Specifically, by clicking on "service credentials" followed by "new credentials" and "add", the username and password required for using the API will become available. Figure 3 illustrates some of the above steps.

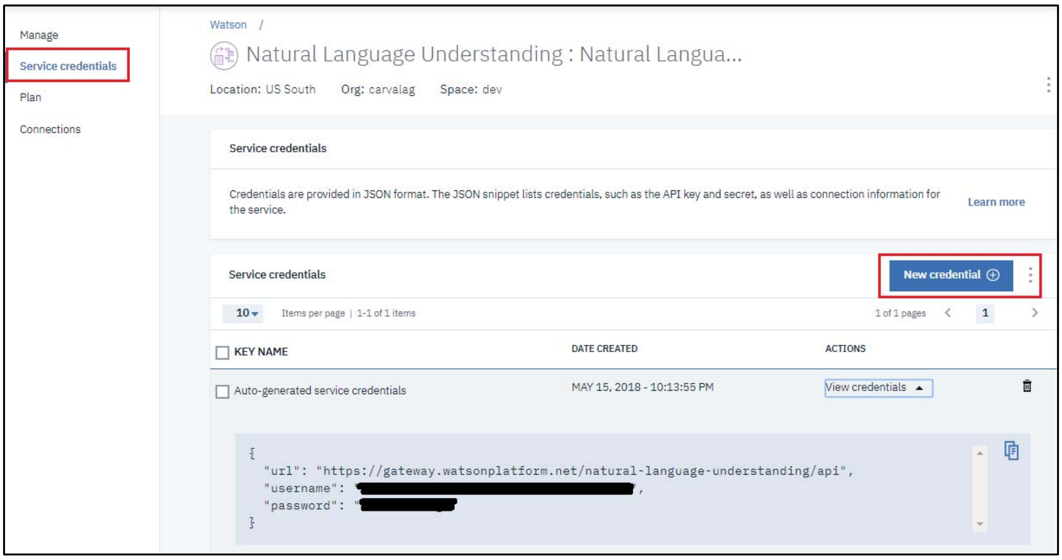


Figure 3. Illustration of the Steps to Obtain the Required Credentials to Use NLU

2.2 Using NLU's API

We can now use NLU's API by first making a REST request, and then interpreting the obtained JSON file. In the following example, we use the same text we previously used with NLU's Web interface. Algorithm 1 shows a script in the programming language R that sends the above text to NLU as a GET operation. Note that the R code does not preprocess the textual data a priori (e.g., it does not perform any tokenization, create n-grams, remove stop words, perform stemming operations, etc.). The first line in Algorithm 1 (see Figure 4) loads the required R library ("httr") to perform the REST request. The following three lines define the base URL, password, and username used in conjunction with the REST request. The rest of the code comprises a simple GET operation using the base URL, username, and password for authentication and three key-value pairs as part of the query. The first key-value pair, which the key "version" defines, determines the API version. The second key-value pair, which the key "text" defines, determines the text that one wants NLU to analyze. Lastly, the key-value pair, which the key "features" defines, determines the analysis that one wants NLU to perform. One can request seven features: "sentiment", "emotion", "keywords", "entities", "categories", "concepts", and "semantic_roles". We focus on the first two features so as to retrieve sentiment and emotion scores.

```
library("httr")
url      <- "https://gateway.watsonplatform.net/natural-language-understanding/api/v1/analyze"
username <- "enter username here"
password <- "enter password here"

#comment: the variable "text" is soft-wrapped, i.e., there is no line breaking
response <- GET(url, query = list(version="2017-02-27",
                 text = "To the Starbucks Corporation, as the wife of a uniformed service member, I
                     want to applaud your pledge to employ refugees. Thank you for the ray of
                     hope in a frightening time, and for demonstrating humanity in the realm of
                     business",
                 features = "sentiment,emotion"),
                 authenticate(username, password))
```

Figure 4. Algorithm 1: Sample R Code that Requests Sentiment/Emotion Scores from NLU

After running Algorithm 1, one obtains a JSON file that contains the requested data. Algorithm 2 (see Figure 5) shows an R script to process the obtained data so as to retrieve sentiment and emotion scores. The first line loads the required package (“jsonlite”) to process the JSON file. The second line retrieves the content of the response to the GET request (i.e., the data in JSON). The next two lines retrieve the sentiment score (0.9954) and the “sadness” score (0.1468). Figure 6 shows the structure of the returned JSON file.

```r
library("jsonlite")
response <- fromJSON(content(response, "text"))

#comment: retrieving sentiment score
response$sentiment$document$score

#comment: retrieving sadness score
response$emotion$document$emotion$sadness
```

Figure 5. Algorithm 2: Sample R Code that Obtains Sentiment/Emotion Scores from NLU's Response

```
{
    "usage": { "text_units": [1],
               "text_characters": [231],
               "features": [2]
    },
    "sentiment": {
        "document": {  "score":[ 0.9954],
                       "label": ["positive"]
        }
    },
    "language": ["en"],
    "emotion": {
        "document": {
            "emotion": {
                "sadness": [0.1468],
                "joy": [0.4209],
                "fear": [0.0855],
                "disgust": [0.2932],
                "anger": [0.1115]
            }
        }
    }
}
```

Figure 6. JSON File Associated with NLU’s Response

Algorithm 2 traverses the tree-like structure of the JSON file to retrieve the data of interest. In particular, one retrieves each subtree that a certain key defines by using the key label after the dollar sign ($) symbol. During the data-collection phase in our case study, we used NLU's API to programmatically obtain sentiment and emotion scores for thousands of social media comments. Note once again that we did not preprocess the data at all before calling NLU's API—we simply submitted to NLU the original Facebook posts without any a priori textual manipulation, which highlights how easy it is to use cloud-based AI technologies (i.e., other than programming skills, one requires no other analytics-related skills). We further discuss this point in Section 4. We further elaborate on how we collected data and applied NLU in Section 3.

3 Case Study: Understand the Implications of a Starbucks Corporate Social Responsibility Effort

In this study, we illustrate how to effectively use NLU, an off-the-shelf AI technology, by considering the case of an analytics project that follows the Cross-Industry Standard Process for Data Mining (CRISP-DM). CRISP-DM defines a traditional data-analytics project in terms of six phases (Shearer, 2000) as we show in Figure 7 and discuss below:

1) Business understanding: this first phase involves deeply understanding the underlying business problem and analytics goals.
2) Data understanding: this phase primarily involves collecting and exploring data.
3) Data preparation: in this phase, one merges and preprocesses the acquired data sets, from potentially different sources, to obtain a unified data set suitable for analysis.
4) Modeling: in this step, one analyzes the data, which includes selecting statistical-modeling techniques.
5) Evaluation: at this stage, one evaluates the results from the performed analyses in light of the business problem, which includes evaluating factors such as a statistical model's accuracy and generality.
6) Deployment: at the end of the CRISP-DM process, one needs to appropriately organize and report the knowledge obtained from the performed analyses. Depending on the underlying requirements, the deployment phase can be as simple as generating a report or as complex as deploying a statistical model that one will use to, for example, make real-time predictions.

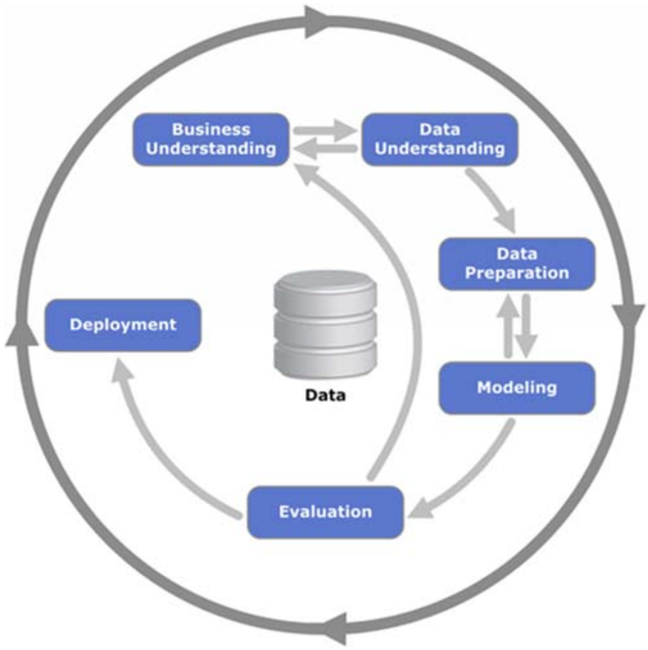


Figure 7. The Six Phases of CRISP-DM

Given the complexity and heterogeneity of the above phases, one may express skepticism about the possibility that a single technology could handle all the data-analytics process. A more appropriate perspective involves considering that different technologies, including AI-based technologies, can help data analysts during specific tasks in specific phases. To illustrate this perspective, we exemplify in this section how to use NLU to assign sentiment and emotion scores to social-media posts during the data-preparation phase in a data-analytics project. The case we consider relates to the pledge Starbucks made on 29 January, 2017, to hire thousands of refugees. We investigated how that pledge affected consumers' and investors' sentiment towards Starbucks. To obtain consumers' sentiment towards Starbucks, we first collected user-generated posts on Starbucks' Facebook page. As a result, we obtained highly unstructured data. In order to prepare the data for analysis, we applied NLU to extract the sentiment and emotions behind each post. In addition to illustrate how one can use off-the-shelf AI technologies during a data-analytics project, we generate insights from our analysis that can help organizations that are contemplating corporate social responsibility strategies.

We organize this section in line with the CRISP-DM process in a sense that, except for the deployment phase, each subsection reflects one or more CRISP-DM phase. First, we provide the appropriate background by elaborating on the business problem and by positioning our work against the relevant literature. Next, we explain the data-collection and -preparation processes we followed. Subsequently, we describe how we analyzed the data. Finally, we evaluate our results.

3.1 Problem Understanding

Recent years have seen a tremendous growth in the number of refugees across the globe. The United Nations High Commissioner for Refugees (UNHCR) (2017) has estimated that, by the end of 2017, internal conflicts, persecution, generalized violence, or human rights violations forcibly displaced a total of 68.5 million individuals. Further, the UNHCR (2017) considered 25.4 million of those 68.5 million individuals as refugees. Unfortunately, these numbers seem to be rising at a fast pace as UNHCR (2017) estimated 2.9 million more displaced persons in 2017 than in 2016 (UNHCR, 2017).

Public and private organizations have reacted to this massive humanitarian crisis in different ways. For example, TripAdvisor committed more than five million dollars (USD) to meeting urgent humanitarian needs in Europe and in the Syria region (TripAdvisor, 2017). Companies such as HP, Google, Microsoft, Intel, J.P. Morgan Chase, and Goldman Sachs announced commitments intended to bolster educational and employment opportunities for refugees in several countries (Jordan, 2016).

Other companies have gone beyond promising financial donations and training. In response to President Donald Trump's Executive Order 13769, which, among other things, banned people from predominantly Muslim countries (Iran, Iraq, Libya, Somalia, Sudan, Syria, and Yemen) from entering the USA (White House, 2017), Starbucks announced on 29 January, 2017, its plan to hire 10,000 refugees globally by 2022. In particular, Howard Schultz, CEO and Chairman of Starbucks at that point in time, stated:

> *There are more than 65 million citizens of the world recognized as refugees by the United Nations, and we are developing plans to hire 10,000 of them over five years in the 75 countries around the world where Starbucks does business. And we will start this effort here in the U.S. by making the initial focus of our hiring efforts on those individuals who have served with U.S. troops as interpreters and support personnel in the various countries where our military has asked for such support.* (Starbucks, 2017)

While consumers often have high regard for companies that offer financial donations (Dean, 2003), Starbucks' hiring announcement prompted both praise and backlash on social media. As we elaborate on throughout this section, our data suggest that, while some applauded Starbucks on its Facebook page for its humanitarian efforts (e.g., "Thank you so much for your stand on helping refugees by offering jobs. I don't usually buy coffee at Starbucks, but make my own Starbucks at home. I plan on making a point of buying more coffee at Starbucks as a way of showing my support."), others have criticized the company for allocating job opportunities to refugees as opposed to hiring, for example, war veterans (e.g., "Why do you hate America? The homeless in America? Our veterans and all others that are unemployed here?").

In Section 3.3, we formally investigate what impact the announcement of pro-refugee hiring policies might have on organizations. We also consider how off-the-shelf AI technologies might help with this task. Focusing on the Starbucks case, we study the impact that its hiring announcement had on consumers' and investors' sentiment towards the company. Clearly, the current refugee crisis has yet to near its end (Esses, Hamilton, & Gaucher, 2017). With our case study, we elucidate the consequences of Starbucks' pledge so that companies considering a similar policy can better understand its implications. At the same time, we illustrate the potential effectiveness of off-the-shelf AI technologies. Specifically, we address the following research questions (RQ):

RQ1: How can off-the-shelf AI technologies help with analyzing comments/opinions collected from social media?

RQ2: How did consumers' sentiment towards Starbucks after the hiring announcement compare to consumers' sentiment before the hiring announcement?

RQ3: Did consumers' emotions shift significantly due to the pro-refugee hiring plan?

RQ4: How did investors' sentiments towards Starbucks after the hiring announcement compare to their sentiments before the hiring announcement?

RQ5: Can potential changes to consumers' sentiments explain any potential changes to investors' sentiments?

Since our case study speaks to several lines of research, we provide some extra background and review the relevant literature in Sections 3.1.1 to 3.1.3.

3.1.1 Starbucks

As we mention above, Starbucks pledged to hire 10,000 refugees on 29 January, 2017. This announcement somewhat differed from other companies' announcements because it did not involve financial donations or allocation of resources for training/education purposes. Instead, it was about offering refugees jobs, a practice that can potentially fuel the rhetoric that immigrants take jobs from a country's current citizens.

With over 24,000 stores in 70 countries, roughly 254,000 employees, and revenues of US$21.31 billion, Starbucks ranked 131st in the 2016 Fortune 500 ranking, third in Fortune's "World's Most Admired Companies" ranking, and number 45 in Fortune's "Change the World" ranking (Fortune, 2017). Furthermore, Starbucks is widely regarded as a leader in using customer-focused social media and social media engagement (Chua & Banerjee, 2013; Gallaugher & Ransbotham, 2010). One needs to consider this latter fact in particular to understand how off-the-shelf AI technologies can help with analyzing social-media data (RQ1) since it means that social media data related to Starbucks will likely be widely available.

Examining social-media data related to Starbucks allows one to understand the impact of a pro-refugee hiring pledge from a firm that operates under a bright public spotlight. Moreover, Starbucks attracts a great number of both admirers and detractors. On one hand, Starbucks often receives vocal criticism from individuals who see the company as a symbol of global corporate sprawl and consumerism. On the other hand, Starbucks has strong corporate social responsibility policies, which we discuss in Section 3.1.2.

3.1.2 Corporate Social Responsibility

One can define corporate social responsibility as a company's commitment to minimizing or eliminating any harmful effects on society and maximizing its long-term beneficial impact (Mohr, Webb, & Harris, 2001). This definition incorporates actions such as obeying laws and ethical norms, treating employees fairly (see the discussion about Foxconn by Ma, Shang, & Wang, 2017), protecting the environment (Hsu & Wang, 2013), and contributing to charities and philanthropic projects (Partovi, 2011).

In its early days, individual firms took corporate social responsibility as a voluntary action. Following legislated responsibilities, many companies now view corporate social responsibility as an investment that might improve their long-term performance (Varadarajan & Menon, 1988). In other words, actions undertaken as corporate social responsibility may be partly altruistic, but they may also serve corporate self-interest. Dean (2003) suggests that corporate social responsibility may result in the following benefits for an organization: 1) creation of goodwill with a community, 2) differentiation of the corporate image and its brands from competitors, 3) greater customer acceptance of price increases, 4) increase in employee and channel member morale, 5) recruitment of new employees, 6) use as a shield against public criticism in times of crisis, 7) increased likelihood to win over skeptical public officials, and 8) increased revenues and profits.

Oftentimes, corporate social responsibility equates to financial donations. The experimental results from Varadarajan and Menon (1988) suggest that unconditional donations to charities never harm a company's image, whereas conditional donations (i.e., donations tied to revenue-producing transactions) might. Our work differs from previous work in that we focus on the case where a company promises jobs to a certain group of individuals (refugees). Such a promise can polarize consumers and, consequently, damage the company's image since job allocation might be seen as a zero-sum situation (e.g., a refugee getting a job might signal that a citizen does not get it), which can foster an "us-against-them" view and lead to the perception that the hiring company favors "foreigners" as opposed to "current citizens" (a comment that we found several times in our data set). As a consequence, the sentiment of consumers towards the hiring company might become more negative. To investigate the influence of Starbucks' refugee hiring pledge on consumers' sentiments, we collected and analyzed comments (textual data) from social media—a practice called social media analytics, which we discuss in Section 3.1.3.

3.1.3 Social Media Analytics

With the advent of social media, companies have considerably strengthened their ability to engage in company-customer dialog, to create mechanisms for customer-customer dialog, and, perhaps more importantly, to monitor and mediate these dialogues. Recognizing the important role social media plays today, academics have contributed a significant number of academic publications on the topic that range from understanding the impact of rewards programs (Mirzaei, Odegaard, & Yan, 2015) to developing strategies for increasing network size in social networks (Ballings, Van den Poel, & Bogaert, 2016).

Obviously, the new communication channels between companies and consumers generate precious and often highly unstructured data (e.g., comments in natural language, pictures, videos, etc.). Social media analytics refers to the task of collecting and analyzing data from social media (Fan & Gordon, 2014). In this paper, we practice social media analytics when collecting and analyzing comments before and after Starbucks made its refugee announcement. In particular, we assign scores to each comment that reflect the sentiment and emotion behind that comment. Thereafter, we measure whether the scores before and after the announcement changed significantly.

One can assign sentiment scores to a text in many different ways. A classic approach called the bag-of-words approach involves simply subtracting the number of negative words from the number of positive words. The higher the resulting value, the more positive the sentiment behind the text. One can use some dictionaries to determine the polarity and intensity of each word in a text (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). This descriptive approach is, however, rather naïve because it ignores syntax. More recently, researchers have proposed several techniques to deal with this issue while trying to mimic the way humans reason about natural language. These techniques belong to a field called natural language processing, which, in turn, many consider as a subfield of the artificial intelligence area (Russell & Norvig, 2016).

3.2 Data Collection and Preprocessing

To answer the research questions we describe in Section 3.1, we first collected data from Starbucks' page on Facebook (https://www.facebook.com/Starbucks). Thereafter, we used NLU to obtain sentiment and emotion data as we describe in Section 2.2. Finally, we collected stock-price data from NASDAQ to answer the last two research questions. In this section, we describe in detail the data-collection processes, which we performed using the programming language/statistical environment R.

3.2.1 Facebook Data

To measure customers' sentiment towards Starbucks, we collected user-generated posts from Starbucks' public page on Facebook. In order to make clear comparisons between sentiments and emotions before and after Starbucks announced its hiring plan, the collected data only include posts from the 15 days directly before and after the announcement on 29 January, 2017. That is, we collected posts that appeared on the public page between 15 January, 2017, and 13 February, 2017. We found this time window wide enough to provide sufficient data for statistical analysis but narrow enough to avoid the influence of secondary factors, such as seasonality, on the posts. We consider all the posts posted on 29 January as part of the before-announcement data.

Facebook posts can provide insights into how consumers feel at specific points in time, which directly aligns with our focus on identifying possible changes in consumers' sentiments and emotions following Starbucks' hiring announcement. One advantage of using Facebook instead of other social media sites, such as Twitter, is that it has lax character limits. In order words, Facebook allows users to express their complete opinions without facing severe restrictions on the number of written characters, an important aspect when estimating sentiment and emotion scores. In addition, a very wide audience uses Facebook (Greenwood, Perrin, & Duggan, 2016), and, according to Gallaugher and Ransbotham's (2010, p. 198) Starbucks case study, "the demographics and behavior patterns of Starbucks' customers align well with social media".

However, we do acknowledge potential limitations associated with using Facebook posts. First, one must assume that Facebook users who post on Starbucks' page on Facebook represent the population of Starbucks consumers. It is virtually impossible to verify whether all users who post on that page actually patronize Starbucks. In our context, some users may have posted due to an emotional response to Starbucks' announcement even if they did not patronize Starbucks. Nevertheless, we can reasonably assume that many of the people who post on Starbucks' page have an interest in Starbucks and patronize their stores. Second, Facebook users might have posted comments unrelated to the hiring announcement (e.g., "I dislike Starbucks coffee"). To detect such statements in our data set, we would have had to go through all the collected posts and make a subjective call on whether or not each post concerned the hiring plan. We decided against making any subjective judgment and, consequently, against removing any data point based on the rationale that potential consumers could post comments unrelated to the hiring plan anytime. Hence, such unrelated posts before and after the announcement likely had a similar average sentiment/emotion, which means they would likely not affect our analysis.

We collected data from Starbucks' Facebook page by using Facebook's Graph API, with which we could gather users' posts on that page. We did not collect comments on the original posts because we realized that several comments involved users' bad-mouthing other users and, thus, did not concern the original pledge issue. In total, our data set comprised 3,308 posts of varying lengths. After removing posts that Starbucks itself wrote, we ended up with 3,296 posts: 411 posts during the 15 days before the announcement, and 2,885 posts during the 15 days after the announcement.

3.2.2 Sentiment and Emotion Data

We next preprocessed the data collected from Facebook using IBM NLU's API as we describe in Section 2.2. The extracted sentiment scores ranged from -1 (purely negative) to 1 (purely positive) with a score of 0 representing a neutral post. The extracted emotion scores ranged from 0 (absent) to 1 (extremely present). We collected scores associated with all five available emotions; namely, sadness, joy, fear, disgust, and anger. This data-preprocessing step effectively answers RQ1 in that it shows how an off-the-shelf AI technology can help one analyze comments/opinions collected from social media.

Unfortunately, the current version of NLU cannot estimate emotion scores when receiving very short sentences as input (e.g., single words). After removing all the observations without sentiment and/or emotion scores due to the same being very short, we end up with 3,189 observations in our data set: 399 posts in the 15 days before the announcement, and 2,790 posts in the 15 days after the announcement.

3.2.3 Stock Prices Data

In addition to Facebook and sentiment/emotion data, we collected Starbucks' stock prices (SBUX) from NASDAQ (NASDAQ, 2017). We took stock prices as a proxy for investors' sentiment towards a company's performance at a given point in time. Doing so concurs with results from behavioral finance studies, which suggest that investors' sentiments do indeed considerably affect stock prices (Baker & Wurgler, 2007). That said, we collected 30 minute-level intraday stock prices, including overnight trading, between 15 January and 13 February, 2017, to gauge investors' sentiment towards the Starbucks brand before and after it announced the hiring plan. We considered all stock prices on 29 January as before-announcement data. Various events, such as Starbucks' hiring plan, can quickly affect investors' sentiment, which a stock price can subsequently reflect. The fluid nature of a stock price makes it a suitable variable to consider when determining whether a significant relationship exists between consumers' and investors' sentiment towards Starbucks (see our RQ5 in Section 3.1).

3.3 Data Analysis

In this section, we discuss how we analyzed the collected data to answer RQ2 to RQ5 (see Section 3.1). In particular, in each subsection, we address a specific research question. We performed the data analysis using the programming language/statistical environment R.

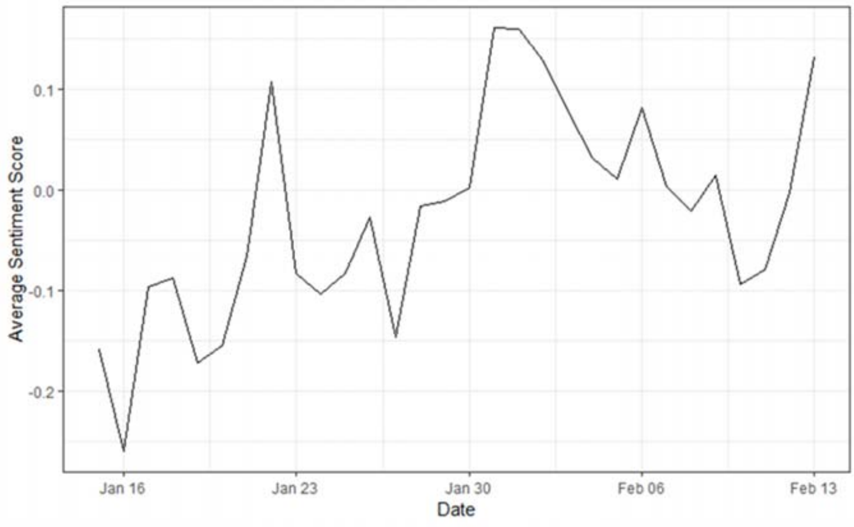


Figure 8. Average Sentiment Score per Day

3.3.1 Analysis of Sentiment Data

We first analyze RQ2; namely, how consumers' sentiment towards Starbucks after the hiring announcement compared to consumers' sentiment before the hiring announcement. Despite the controversial nature of Starbucks' hiring plan, we found that the average consumer sentiment towards the company increased after the pledge. While the average sentiment score noticeably fluctuated from day to day, we found a peak during the days immediately following the announcement as one can see in Figure 8. However, it appears that the average sentiment score returned to pre-announcement levels relatively quickly. One can make a similar statement about the number of posts (see Figure 9).

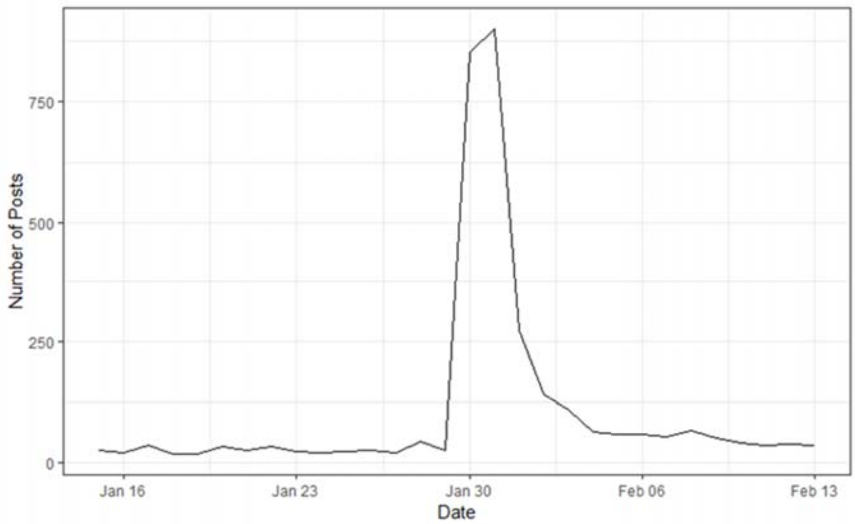


Figure 9. Number of Posts on Starbucks' Facebook Page

Given the results from our exploratory plot in Figure 8, we next performed a one-tailed t-test. Specifically, the null hypothesis is that the average sentiment score before and after the announcement were the same, and the alternative hypothesis is that the average sentiment score after the announcement was greater than before the announcement. The p-value resulting from this statistical test was very small (p-value $< 10^{-7}$, $t = -5.3595$, $df = 546.76$), which provides statistical support that the average sentiment score after the announcement (0.08) was greater than the average score before the announcement (-0.08). Consequently, we conclude that, from a consumer-sentiment perspective, potential consumers on average received well the pro-refugee hiring policy that Starbucks announced.

We also conducted a time-series analysis in which we looked for structural breaks/changes as another way to show that the hiring announcement did actually influence consumers' sentiments towards Starbucks. Specifically, after sorting the Facebook posts by the date and time people posted them, we next built two simple linear regression models with the sentiment score received by posts (*SentimentScore*) as the dependent variable and the position of the posts in the sorted array of posts (*t*) as the independent variable. The first regression model contained the posts numbered 1 through 399 (pre-announcement data), whereas the second regression model contained the posts numbered 400 to 3189 (post-announcement data). Formally:

$$SentimentScore_t = \beta_0 + \beta_1 * t + \epsilon_t \text{ for } t \in \{1, 2, \dots, 399\} \quad \text{(Model 1)}$$

$$SentimentScore_t = \delta_0 + \delta_1 * t + \epsilon_t, \text{ for } t \in \{400, 401, \dots, 3189\} \quad \text{(Model 2)}$$

We next ran a Chow test to determine whether the coefficients of the linear models were the same (null hypothesis) or not (alternative hypothesis); that is, whether $\beta_0 = \delta_0$ and $\beta_1 = \delta_1$. The results from this test (F statistic = 4.86, p-value = 0.0078) shows that we had enough statistical evidence to say that the coefficients were not the same and, hence, the day of the announcement constituted a structural change in the sentiment-score time series.

Note that the above results are based on average sentiment scores, which can obscure a crucial point regarding Starbucks' hiring announcement; namely, that it created a great deal of polarization among potential consumers. Figure 10 illustrates as much by plotting the number of positive posts (sentiment score greater than zero) against the number of negative posts (sentiment score less than zero) per day. Although the number of positive posts was generally greater than the number of negative posts right after the announcement, we can see a peak in the number of negative posts around 30 January, which shows that by no means was every single potential consumer pleased with Starbucks' pledge to hire refugees. Also note how all the above claims heavily depend on how NLU assigns sentiment scores, which supports our primary goal in this section to illustrate how one can use an off-the-shelf AI technology in data-analytics projects (RQ1).

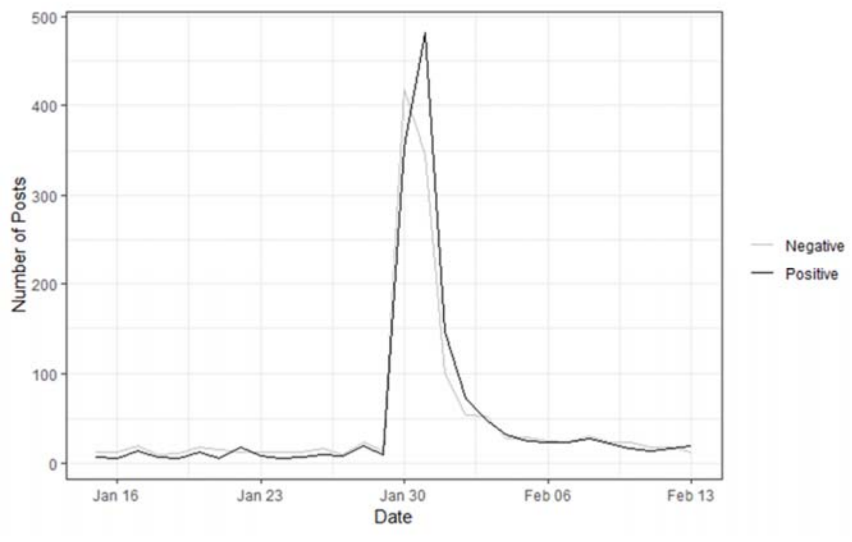


Figure 10. Number of Posts on Starbucks' Facebook Page by Sentiment

3.3.2 Analysis of Emotion Data

Although the average sentiment among consumers increased following Starbucks' press release, specific emotions might have changed significantly due to its hiring plan. RQ3 (i.e., whether consumers' emotions shift significantly due to the pro-refugee hiring plan) addresses this issue. Our preliminary exploratory analysis did not prove helpful in providing a picture regarding changes in emotions. As such, to address our RQ3, we performed a two-tailed t-test for each one of the five emotions (sadness, joy, fear, disgust, and anger) to measure whether any change in average emotion score from before to after the announcement was statistically significant. The null hypothesis was that the average emotion before the announcement was the same after the announcement, whereas the alternative hypothesis was that the average emotion before the announcement was not the same as after the announcement. Table 1 summarizes the results of our tests.

Table 1. Results of the Statistical Tests Regarding Emotion Scores

Emotion	Mean score before pledge	Mean score after pledge	T test			Chow test	
			P value	T value	DF	P value	F statistic
Sadness	0.3087	0.2870	0.085	1.726	504.47	0.2	1.608
Joy	0.3651	0.3522	0.427	0.795	513.87	0.0004	7.8431
Fear	0.0869	0.0757	0.032	2.142	512.16	0.0456	3.09
Disgust	0.1251	0.2077	$< 10^{-15}$	-10.263	616.84	$< 10^{-15}$	67.6
Anger	0.1658	0.1386	0.002	3.11	481.4	0.0037	5.613

Table 1 shows that, for the t-test results, negative emotions such as fear and anger had significantly greater before-announcement scores than after-announcement scores. To a certain degree, these results support our previous sentiment analysis in that potential Starbucks consumers generally received the hiring pledge well. However, the increase in disgust paints a different picture. We cannot easily interpret this result given the textual data we have. For example, the increase in disgust might have resulted from Starbucks' hiring policy or because some consumers felt uneasy with the underlying refugee crisis. The Chow tests did not provide entirely conclusive results as well. Specifically, they showed that structural changes in the emotion-score time series for joy, fear, disgust, and anger but not for sadness. However, both tests found significant changes in fear, disgust, and anger.

Despite the fact that we did not find entirely conclusive results, we nonetheless take them as further evidence about how polarizing Starbucks' announcement was. Consumers definitely expressed mixed feelings as different changes in emotion scores illustrate. We discuss potential business implications that result from such changes in Section 3.4. Once again, we highlight that we could conduct the analysis above due to the emotion scores that an off-the-shelf AI technology assigned. As such, we illustrate how one can use such a technology to help preprocess and analyze data (RQ1).

3.3.3 Analysis of Investors' Sentiment

We next analyzed Starbucks' stock prices before and after the hiring pledge. Of course, stockholders represent one group of people who pay particular attention to Starbucks' announcements. Since stockholders buy and sell stocks based on their expectations of a company, we can reasonably assume that, to a certain degree, a stock price reflects investors' sentiment towards the underlying company (Baker & Wurgler, 2007). With RQ4, we investigate changes in investor sentiment levels before and after the hiring announcement.

To gain some intuition, we first plotted Starbucks' closing stock price over the 30 days that our data set covered (see Figure 11). Interestingly, one can see a sharp decline in closing prices right after the hiring announcement, which suggests that investors did not approve Starbucks' plan to hire refugees. From a business perspective, one can understand this decline since one cannot easily contemplate a priori how such a plan might increase Starbucks' profitability and/or value as a brand. Further, the decline in investors' sentiment goes against the average increase in consumers' sentiment towards Starbucks. We return to this point in Section 3.3.4.

Given the results from our exploratory plot in Figure 11, we next performed a one-tailed t-test. Specifically, the null hypothesis was that the average stock price before the announcement was the same after the announcement, and the alternative hypothesis was that the average stock price after the announcement was less than before the announcement. This statistical test produced a very small p-value (p-value$< 10^{-16}$, $t = 37.648$, $df = 446.44$), which provides statistical support that the average price after the announcement (55.46) was less than the average price before the announcement (57.41). The Chow test also corroborated the above result ($F = 759.91$, p-value $<10^{-16}$) by showing that the announcement day created a structural change in the stock-price time series. Consequently, we conclude that, from a stock-price perspective, investors on average badly received the pro-refugee policy that Starbucks announced.

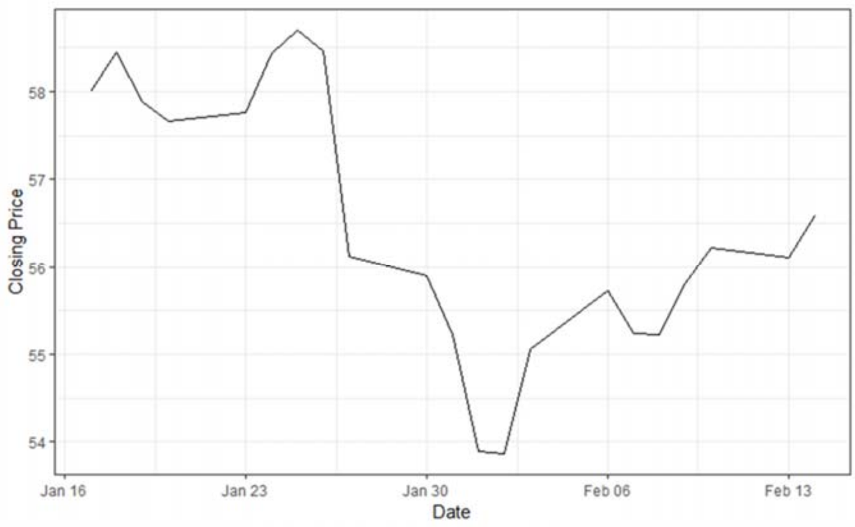


Figure 11. Starbucks' Closing Stock Price in Dollars: Note the Range of the Y-Axis: [53.6, 58.9]

One could argue that the market's overall performance could have driven the above results as opposed to Starbucks' hiring announcement. We investigated this issue by building a regression model with Starbucks' stock price as the dependent variable (*StockPrice*) and the following predictors: 1) NASDAQ Composite Index values (*NasdaqComp*) and 2) a dummy variable that controlled for whether a certain day comes before the hiring announcement (*AnnouncementBefore*). The NASDAQ Composite Index is the market capitalization-weighted index of approximately 3,000 common equities listed on the NASDAQ stock exchange. Similar to how we collected Starbucks' stock prices, we collected 30 minute-level intraday values of the NASDAQ Composite Index between 15 January, 2017, and 13 February, 2017. We merged the stock prices and the index values by date and time and removed observations that had missing values (e.g., stock prices that resulted from after-hours trading that had no associated index values). The dummy variable had the value 0 for trading days before the announcement day and 1 otherwise. Such a modeling choice constitutes common practice in analyzing before-after data (Gelman, 2004). Thus, our regression model, whose estimates coefficients Table 2 displays, had the form:

$$StockPrice_t = \beta_0 + \beta_1 * NasdaqComp_t + \beta_2 * AnnouncementBefore_t + \epsilon_t \qquad \text{(Model 3)}$$

Table 2. Regression Model that Investigates the Influence of Market Performance on Starbucks' Stock Price

Coefficients	Estimate	Std. error	T value	P value
Intercept (β_0)	66.131	1.634	40.464	$<10^{-16}$
NasdaqComp (β_1)	-0.0019	0.0003	-6.531	$<10^{-11}$
AnnouncementBefore (β_2)	-1.8609	0.0403	-46.166	$<10^{-16}$
Residual standard error: 0.7005 on 3186 degrees of freedom Multiple R^2: 0.465 F-statistic: 1384 on 2 and 3186 degrees of freedom (p-value: $< 10^{-16}$)				

Table 2 shows that, even when we controlled for NASDAQ composite index values, the coefficient representing Starbucks' announcement of the pro-refugee hiring policy (β_2) was significant. Furthermore, its negative value (-1.8609) further corroborates our conclusion that investors' sentiment towards Starbucks significantly declined after the hiring pledge. In other words, holding everything else constant, the before-announcement stock price was on average US$1.86 higher than the after-announcement stock price. Both the r-squared and the F-statistic measures show that our model fit the underlying data well.

3.3.4 Analysis of the Relationship between Investors' and Consumers' Sentiments

It seems from the previous analyses that, on average, consumers and investors reacted differently to Starbucks' hiring plan. In particular, the average consumer sentiment tended to increase after Starbucks' announcement, whereas investor sentiment tended to decrease. Therefore, we formally examined whether investors' sentiment towards Starbucks correlated with consumers' sentiment (i.e., RQ5).

We started by merging sentiment data with stock-price data by date and time. Recall that we collected 30 minute-level intraday stock prices. To match sentiment scores with prices, we looked at the time the underlying Facebook post appeared and matched it with the immediately ahead stock price. Hence, each observation in our merged data set represents a single Facebook post and its sentiment score alongside an associated stock price. We used such a merged data set to build a regression model in order to examine the relationship between consumers' and investors' sentiments towards Starbucks. Specifically, we used stock price (*StockPrice*) as the dependent variable in our regression model, which also had three explanatory variables: 1) the sentiment score of a post (*SentimentScore*), 2) a dummy variable that controlled for whether a certain post came before the hiring announcement (*AnnouncementBefore*), and 3) the date of the Facebook post (*Date*). As in our previous regression analysis, the dummy variable receives the value 0 for days before the announcement day, and 1 otherwise. Since the number of posts and, consequently, the number of sentiment scores greatly vary for different days (see Figure 9), we then develop the following random-effect model:

$$StockPrice_{ij} = \beta_0 + \beta_1 * SentimentScore_{ij} + \beta_2 * AnnouncementBefore_j + Date_j + \epsilon_{ij} \qquad \text{(Model 4)}$$

In the above equation, *StockPrice*$_{ij}$ and *SentimentScore*$_{ij}$ are, respectively, the *i*th price/score on day *j*, whereas *Date*$_j$ is the day-specific random effect. We found that including a random effect yielded a higher-quality model with lower Akaike information criterion (AIC) (2389.307). Specifically, a fixed-effect model with *SentimentScore* as the only predictor had an AIC of 8745.943, whereas a fixed-effect model with both *SentimentScore* and *AnnouncementBefore* as predictors had an AIC of 6817.902. Table 3 summarizes the estimated coefficients of our model. We estimated p-values based on Satterthwaite's method (Kuznetsova, Brockhoff, & Christensen, 2017). We derived conditional and marginal coefficient of determinations based on the techniques that Nakagawa and Schielzeth (2013) propose.

Table 3. Regression Model that Investigates the Relationship between Consumers' and Investors' Sentiments towards Starbucks

Random effects		
Groups	**Variance**	**Std. dev.**
Date	0.6980	0.8354
Residual (ϵ)	0.1168	0.3417

Fixed effects						
Coefficients	**Estimate**	**Std. error**	**T value**	**p value**	**Confidence intervals**	
					2.5%	**97.5%**
Intercept (β_0)	55.455	0.21598	256.757	$<10^{-15}$	57.07	57.92
SentimentScore (β_1)	-0.0111	0.01011	-1.104	0.27	-0.03	0.009
AnnouncementBefore (β_2)	-2.0404	0.30576	-6.673	$<10^{-6}$	-2.64	-1.44

ICC (Date): 0.8567
Marginal R^2: 0.3592
Conditional R^2: 0.9081

First, we found that the hiring announcement had a significant influence on Starbucks' stock price given the statistical significance of the coefficient associated with *AnnouncementBefore*, which corroborates our previous analysis from Section 3.3.3 regarding that influence. More interestingly, we did not have enough statistical evidence to suggest that consumers' sentiment drove investors' sentiment towards Starbucks. One can take this result as a word of caution against the predictability of stock prices based on activities on social media, a popular research topic at present (Tumarkin & Whitelaw, 2001; Chen, De, Hu, & Hwang, 2014; Jerdack, Dauletbek, Divine, Hult, & Carvalho, 2018). We return to this issue in Section 3.4. Finally, we highlight that both r-squared measures show that our model fit our data well.

3.4 Evaluation of the Results

Several companies have reacted in different ways to the current refugee crises. In our case study, we investigated the impact that Starbuck's announcement about refugee hiring policies had on potential consumers' and investors' sentiments towards the company. We found that the average sentiment of potential consumers towards Starbucks went up after the hiring announcement. However, not everyone shared this positive feeling: the announcement greatly polarized potential consumers, which led some of them to despise the hiring pledge. We also found that the hiring announcement negatively affected Starbucks' stock prices, which we took as a proxy for investor sentiment. Finally, we found no evidence that consumers' sentiment affected investors' sentiment.

Given the above results, the following question arises: does it benefit a company to (promise to) hire refugees as part of its corporate-social-responsibility policies? If we disregard the humanitarian and welfare aspects that such a policy would bring and focus purely on immediate, short-term outcomes as measured by consumer sentiment and stock valuation, then we conclude that such a policy might cause more harm than good to a company. For example, the increase in the number of negative comments about Starbucks (see Figure 10) and the us-against-them rhetoric present in many of the negative posts might push some consumers towards boycotting the company. As Simon (2011, p. 145) suggests:

> *They [consumers] have increasingly expressed their ideas about everything from local affairs to foreign relations at the point of purchase—through in this case, not buying a widely recognized product to gain a say in the larger distribution of social power.*

Indeed, we found that the word "boycott" 263 appeared times across the 3,189 comments in our data set.

We hypothesize that such a polarization resulted because, among other reasons, some potential Starbucks consumers might have considered job allocation as a zero-sum game, which means that they interpreted allocating a job to a refugee as not allocating it to a current citizen. Looking at our data set, we found some support for this hypothesis. In particular, we found that consumers commonly complained that Starbucks should hire war veterans as opposed to refugees. Specifically, we found 424 mentions to veterans among the 3,189 comments in our final data set.

We also need to consider not only the overall change in consumers' sentiment brought by the hiring announcement but also specific changes in emotions. With help from the AI technology we used in our analysis, we found that the feeling of disgust increased after Starbucks' announcement. Researchers have often linked an increase in disgust to abandoning regular habits (Han, Lerner, & Zeckhauser, 2012), which could lead some consumers to stop visiting Starbucks' stores altogether. Indeed, researchers have suggested that disgust constitutes a key clue for better understanding boycott motivations and behaviors (Braunsberger & Buckler, 2011).

Besides polarizing potential consumers, the hiring pledge made Starbucks investors more pessimistic since its stock price declined after the underlying announcement. However, surprisingly, we did not have enough statistical evidence to suggest that consumers' sentiment drove investors' sentiment towards Starbucks. The relevant literature reports conflicting results on the predictability of stock returns based on Internet activity. For example, Tumarkin and Whitelaw (2001) found no association between investors' opinions posted on Internet message boards and stock returns. Chen et al. (2014), on the other hand, found a positive association. Given the above conflicting results, our result in Subsection 3.3.4 should caution one from any attempt to explain stock returns based on Internet and, in particular, social media activity.

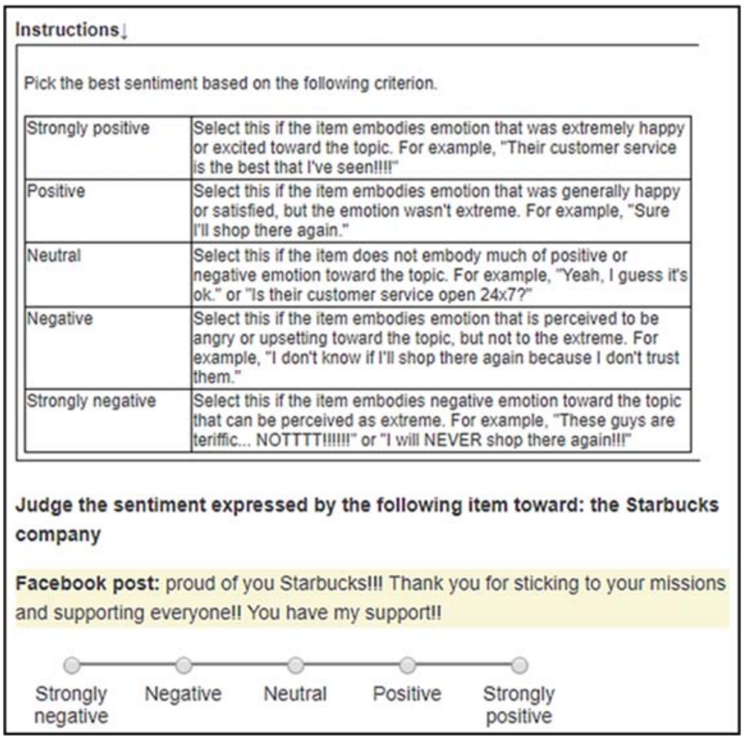
If hiring refugees might not bring short-term benefits to a company, then what would be an ideal corporate social responsibility policy? Di Giuli and Kostovetsky (2014) found no evidence that organizations recover expenditures related to corporate social responsibility policies through increased sales. The authors also found that an increase in an organization's corporate social responsibility ratings were associated with negative future stock returns and declines in returns on assets (ROA). As a consequence, "any benefits to stakeholders from social responsibility come at the direct expense of firm value" (Di Giuli & Kostovetsky, 2014, p. 158). Clearly, similar to ours, such an analysis only takes the perspective of stakeholders and consumers into account. Hence, follow-up work could try to understand the impact that announcing and eventually hiring refugees has on employee-related aspects of an organization, such as employee satisfaction and workforce diversity.

On a final note, we note a potential limitation with the data in our study. NLU, and virtually any AI technology based on natural language processing, might struggle when assigning sentiment/emotion scores to shorts texts. For example, consider the text "Good luck Starbucks!" and its sentiment score of 0.85 according to NLU. In our context, one cannot easily determine the text's meaning. For example, it might mean "Good luck Starbucks with this foolish plan" whose sentiment score equals -0.79. If, on the other hand, the meaning of that text is "Good luck Starbucks with this amazing plan", then the sentiment score becomes 0.9. We do not believe this issue influenced our analysis given that our data set contained relatively long posts (205.63 characters on average).

3.4.1 Evaluating the Accuracy of NLU

The results we report in this section clearly depend on the way the underlying off-the-shelf AI technology assigned sentiment scores to the Facebook posts. Thus, when using similar off-the-shelf technologies, one needs to question how accurate they are. In our context, we note that it is rather troublesome to even quantify accuracy. For example, one can assume that one can measure accuracy by how close sentiment scores that AI technologies assign resemble sentiment scores from humans. But, since sentiment scores are rather subjective, humans might not even agree on what score to assign to a certain text, which explains why researchers commonly request many individuals to label each observation in a data set (Wang, Ipeirotis, & Provost, 2017). At the end, one should use some measure of central tendency to aggregate individual judgments.

We follow a similar approach to evaluate NLU's accuracy. Specifically, we hired workers from the crowdsourcing platform Amazon Mechanical Turk (AMT) to label each Facebook post in our data set. For each Facebook post, we asked five workers to classify the post as strongly negative, negative, neutral, positive, or strongly positive. AMT automatically assigns scores to each possible label (-2, -1, 0, 1, and 2, respectively). The final score of a post is simply the average of the five scores that the workers implicitly reported. Although researchers have published different guidelines on how to select a crowd size (Carvalho, Dimitrov, & Larson, 2015, 2016), we selected five workers per task because AMT currently suggests doing so. We constrained the population of workers to individuals who lived in the United States of America and who had a "master's" qualification (i.e., highly rated workers). Following the AMT's suggestion, we paid US$0.02 for each labeled post. Figure 12 shows an example of the graphical interface.


Instructions

Pick the best sentiment based on the following criterion.

Strongly positive	Select this if the item embodies emotion that was extremely happy or excited toward the topic. For example, "Their customer service is the best that I've seen!!!!"
Positive	Select this if the item embodies emotion that was generally happy or satisfied, but the emotion wasn't extreme. For example, "Sure I'll shop there again."
Neutral	Select this if the item does not embody much of positive or negative emotion toward the topic. For example, "Yeah, I guess it's ok." or "Is their customer service open 24x7?"
Negative	Select this if the item embodies emotion that is perceived to be angry or upsetting toward the topic, but not to the extreme. For example, "I don't know if I'll shop there again because I don't trust them."
Strongly negative	Select this if the item embodies negative emotion toward the topic that can be perceived as extreme. For example, "These guys are teriffic... NOTTTT!!!!!!" or "I will NEVER shop there again!!!"

Judge the sentiment expressed by the following item toward: the Starbucks company

Facebook post: proud of you Starbucks!!! Thank you for sticking to your missions and supporting everyone!! You have my support!!

Strongly negative | Negative | Neutral | Positive | Strongly positive

Figure 22. Graphical Interface of the AMT Task to Label Facebook Posts

After labeling the posts, we could measure NLU's accuracy. We chose the bag-of-word approach, which we briefly discuss in Section 3.1.3, as a natural baseline. We used Hu and Liu's (2004) popular dictionary of positive and negative words. We faced one issue with the difference between the range of AMT scores [-2, 2], NLU scores [-1, 1], and the unbounded scores for the bag-of-words approach. We solved this issue by transforming our problem into a classification problem. Specifically, we relabeled all AMT, NLU, and bag-of-words scores as "negative", "neutral, and "positive" whenever they were, respectively, less than, equal to, and greater than zero. From doing so, we could calculate the overall accuracy of each technique (i.e., how often it agrees with the AMT labels). We found that NLU was accurate 72.64 percent of time, whereas the bag-of-words approach was accurate only 49.77 percent of time. The staggering 22.87 percent difference highlights how accurate off-the-shelf AI technologies can be and how naïve the very popular bag-of-words approach actually is.

3.4.2 Robustness Check: Date Range

In our analysis, we considered a total of 15 days before and after Starbucks made its refugee announcement. One might argue that the date range might have influenced our results. We note that increasing the number of days before and after the announcement day leads to more data points but also adds extra confounds. For example, the posts from up to 30 days before the announcement day included holiday posts, which tend to be positive. Ideally, the data set should include as few days as possible so as to remove the possibility of confounds. However, less days means less data points, and, consequently, some results might no longer be statistically significant. When collecting data from 10 days before and after the announcement day, we found that almost all the sentiment and stock-price results remained qualitatively the same, whereas the changes in emotion scores regarding fear was no longer statistically significant. The above results also remained true when we reduced the date range to five days before and after the announcement day, except that the changes in the emotion scores related to anger and fear were no longer statistically significant. Since the major issue here involves analyzing changes in emotions and we already conclude in Section 3.3.2 that our results are not entirely conclusive (see Subsection 3.3.2), we then take the above discussion as indicating that the date range did not qualitatively affect our results when it comes to consumer and investor sentiments and the relationship between them. We report the results from our robustness checks in the appendix.

4 Conclusion

Artificial intelligence has received a great deal of attention in recent years. The promise that computers can somehow simulate human-thought processes in order to tackle complex computational/analytics problems is certainly appealing but potentially misleading. In particular, it might be easy, for example, for a person uneducated on AI technologies to believe that such technologies can miraculously and almost effortlessly derive precious insights from data. In this paper, we illustrate how AI technologies do not automatically solve an entire data-analytics project but that they can have great value during different project stages. Specifically, focusing on the CRISP-DM methodology, we show how one can use an off-the-shelf AI service, called Natural Language Understanding, to preprocess textual data by assigning sentiment and emotion scores to social media posts. Although we focus on the CRISP-DM methodology's preprocessing phase, one could certainly apply different AI technologies during other CRISP-DM phases. For example, one could use IBM Watson Analytics (O'Leary, 2017) to automatically build and evaluate statistical models during CRISP-DM's modeling and evaluation phases.

Throughout our case study, we explain how sentiment and emotion scores that NLU assigned supported our analysis. Our focus on analyzing textual data has much current relevance given the growing interest in text mining. Nonetheless, we note that one could similarly use off-the-shelf AI technologies when, for example, transcribing audio or recognizing faces in images.

One might ask why researchers and/or companies should use off-the-shelf AI technologies in practice. We note that building expertise in certain AI areas, such as image recognition and text mining, might be costly both timewise and moneywise. Unlike IT giants such as IBM, Microsoft, Google, and Amazon, smaller and/or non-IT companies might not be able to afford to hire highly qualified professionals with expertise in very specific AI topics. In these cases one may find it appealing to use "black-box" AI technologies such as IBM NLU that require only API calls (see Algorithm 1 and 2) and no expertise in the subject matter. The term "black box" is often pejorative and, for our purposes, refers to the fact that it is unclear how a certain technology operates (e.g., one might not know which statistical techniques a technology uses in practice). This uncertainty factor might deter one from using off-the-shelf AI technologies. We nonetheless argue that developers face the same uncertainty issues when developing in-house solutions. For example, when relying on popular machine learning libraries from programming languages such as Python and R and/or statistical environments such as SPSS and SAS, developers place their trust on whoever developed the libraries/software. A positive aspect behind using cloud-based, off-the-shelf AI technologies is that prestigious IT companies usually offer these services, which one can consider as a proxy for good quality.

Besides the potential for greatly speeding up the development of analytics-based products and/or research efforts, one can run off-the-shelf AI technologies relatively cheaply as we discuss in Section 2.1.

The small cost coupled with the easiness of use and high accuracy (see Section 3.4.1) will likely popularize off-the-shelf AI technologies in the near future.

In this paper, we demonstrate how to use off-the-shelf AI technologies in a data-analytics project, and we strongly believe the results from our case study have value to organizations contemplating different corporate social responsibility strategies. Specifically, our results suggest that, although being a noble intention, promising/offering jobs to refugees might harm a company financially in the short term.

References

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. *The Journal of Economic Perspectives, 21*(2), 129-151.

Ballings, M., Van den Poel, D., & Bogaert, M. (2016). Social media optimization: Identifying an optimal strategy for increasing network size on Facebook. *Omega, 59*, 15-25.

Braunsberger, K., & Buckler, B. (2011). What motivates consumers to participate in boycotts: Lessons from the ongoing Canadian seafood boycott. *Journal of Business Research, 64*(1), 96-102.

Carvalho, A., Dimitrov, S., & Larson, K. (2016). How many crowdsourced workers should a requester hire? *Annals of Mathematics and Artificial Intelligence*, *78*(1), 45-72.

Carvalho, A., Dimitrov, S., & Larson, K. (2015). A study on the influence of the number of MTurkers on the quality of the aggregate output. In N. Bulling (Ed.), *Multi-agent systems* (pp. 285-300). Berlin: Springer.

Chen, H., De, P., Hu, Y., & Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies, 27*(5), 1367-1403.

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51-89.

Chua, A. Y., & Banerjee, S. (2013). Customer knowledge management via social media: The case of Starbucks. *Journal of Knowledge Management, 17*(2), 237-249.

Dean, D. H. (2003). Consumer perception of corporate donations effects of company reputation for social responsibility and type of donation. *Journal of Advertising, 32*(4), 91-102.

Di Giuli, A., & Kostovetsky, L. (2014). Are red or blue companies more likely to go green? Politics and corporate social responsibility. *Journal of Financial Economics, 111*(1), 158-180.

Esses, V. M., Hamilton, L. K., & Gaucher, D. (2017). The global refugee crisis: Empirical evidence and policy implications for improving public attitudes and facilitating refugee resettlement. *Social Issues and Policy Review, 11*(1), 78-123.

Fan, W., & Gordon, M. D. (2014). The power of social media analytics. *Communications of the ACM, 57*, 74-81.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., Lally, A., Murdock, J. W., Nyberg, E., Prager, J., Schlaefer, N., & Welty, C. (2010). Building Watson: An overview of the DeepQA Project. *AI Magazine*, *31*(3), 59-79.

Ferrucci, D., Levas, A., Bagchi, S., Gondek, D., & Mueller, E. T. (2013). Watson: Beyond Jeopardy! *Artificial Intelligence*, *199*, 93-105.

Fortune (2017). *Starbucks (SBUX) stock price, financials and news*. Retrieved from http://fortune.com/fortune500/starbucks/

Gallaugher, J., & Ransbotham, S. (2010). Social media and customer dialog management at Starbucks. *MIS Quarterly Executive, 9*(4), 197-212.

Gelman, A. (2004). Treatment effects in before-after data. In X. L. Meng, & A. Gelman (Eds.), *Applied Bayesian modeling and causal inference from incomplete-data perspectives* (pp. 191-198). London, UK: Wiley.

Greenwood, S., Perrin, A., & Duggan, M. (2016). Social media update 2016. *Pew Research Center.* Retrieved from http://www.pewinternet.org/2016/11/11/social-media-update-2016/

Han, S., Lerner, J. S., & Zeckhauser, R. (2012). The disgust-promotes-disposal effect. *Journal of Risk and Uncertainty, 44*(2), 101-113.

Hsu, A. W.-H., & Wang, T. (2013). Does the market value corporate response to climate change? *Omega, 41*(2), 195-206.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 168-177)

IBM. (2015). *IBM acquires AlchemyAPI, enhancing Watson's deep learning capabilities.* Retrieved from https://www-03.ibm.com/press/us/en/pressrelease/46205.wss

Jerdack, N., Dauletbek, A., Divine, M., Hult, M., & Carvalho, A. (2018). Understanding what drives Bitcoin trading activities. In *Proceedings of the Annual Meeting of the Decision Sciences Institute.*

Jordan, M. (2016). Refugee crisis spurs companies to act. *The Wall Street Journal*. Retrieved from https://www.wsj.com/articles/refugee-crisis-spurs-companies-to-act-1474344062

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1-26.

Ma, P., Shang, J., & Wang, H. (2017). Enhancing corporate social responsibility: Contract design under information asymmetry. *Omega, 67*, 19-30.

Mirzaei, F. H., Odegaard, F., & Yan, X. (2015). User reward programs in online social media. *Omega, 57*, 123-144.

Mohr, L. A., Webb, D. J., & Harris, K. E. (2001). Do consumers expect companies to be socially responsible? The impact of corporate social responsibility on buying behavior. *Journal of Consumer Affairs, 35*(1), 45-72.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133-142.

NASDAQ (2017). *Starbucks Corporation common stock historical stock prices.* Retrieved from http://www.nasdaq.com/symbol/sbux/historical

NLU (2018). *Natural language understanding demo*. Retrieved from https://natural-language-understanding-demo.ng.bluemix.net/

O'Leary, D. (2017). Emerging white-collar robotics: The case of Watson analytics. *IEEE Intelligent Systems, 32*(2), 63-67.

Partovi, F. Y. (2011). Corporate philanthropic selection using data envelopment analysis. *Omega, 39*(5), 522-527.

Russell, S. J., & Norvig, P. (2010). *Artificial intelligence: A modern approach*. Upper Saddle River, NJ: Pearson.

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing, 5*(4), 13-22.

Simon, B. (2011). Not going to Starbucks: Boycotts and the out-scouring of politics in the branded world. *Journal of Consumer Culture, 11*(2), 145-167.

Starbucks. (2017). *Message from Howard Schultz to Starbucks partners: Living our values in uncertain times.* Retrieved from https://news.starbucks.com/news/living-our-values-in-uncertain-times

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37*(2), 267-307.

TripAdvisor. (2017). *TripAdvisor announces $5 Million multi-year commitment to refugee crisis.* Retrieved from https://globenewswire.com/news-release/2016/09/20/873073/0/en/TripAdvisor-Announces-5-Million-Multi-Year-Commitment-to-Refugee-Crisis.html

Tumarkin, R., & Whitelaw, R. F. (2001). News or noise? Internet postings and stock prices. *Financial Analysts Journal, 57*(3), 41-51.

UNHCR. (2017). *Global trends: Forced displacement in 2017.* Retrieved from http://www.unhcr.org/globaltrends2017/

Varadarajan, P. R., & Menon, A. (1988). Cause-related marketing: A coalignment of marketing strategy and corporate philanthropy. *The Journal of Marketing*, *52*(3), 58-74.

Wang, J., Ipeirotis, P. G., & Provost, F. (2017). Cost-effective quality assurance in crowd labeling. *Information Systems Research*, *28*(1), 137-158.

White House (2017). *Executive order protecting the nation from foreign terrorist entry into the United States.* Retrieved from https://www.whitehouse.gov/presidential-actions/executive-order-protecting-nation-foreign-terrorist-entry-united-states/

Appendix A: Replication of Previous Analyses for Posts from 10 Days Before and After the Pledge Day

Table A1. Replication of the Main Analyses We Report in Sections 3.3.1 to 3.3.3

Analysis	T test			Chow test	
	P value	T value	DF	P value	F statistic
Sentiment	$< 10^{-3}$	-3.73	345.43	0.40	0.913
Emotion-sadness	0.504	0.67	333.6	0.697	0.361
Emotion-joy	0.112	1.59	333.76	$< 10^{-3}$	8.458
Emotion-fear	0.445	0.76	339.59	0.631	0.460
Emotion-disgust	$< 10^{-15}$	-8.59	378.78	$< 10^{-15}$	43.799
Emotion-anger	0.003	2.98	316.15	< 0.01	6.400
Stock prices	$< 10^{-15}$	26.36	295.64	$< 10^{-15}$	241.59

Table A2. Replication of the Main Analysis We Report in Section 3.3.4

Random effects						
Groups			**Variance**		**Std. dev.**	
Date			0.7284		0.8535	
Residual (ϵ)			0.1230		0.3508	
Fixed effects						
Coefficients	**Estimate**	**Std. error**	**T value**	**P value**	**Confidence intervals**	
					2.5%	**97.5%**
Intercept (β_0)	57.277	0.271	211.54	$<10^{-15}$	56.748	
SentimentScore (β_1)	-0.014	0.011	-1.325	0.185	-0.036	
AnnouncementBefore (β_2)	2.173	0.382	-5.68	$<10^{-4}$	-2.919	
ICC (Date): 0.8555 Marginal R^2: 0.327 Conditional R^2: 0.903						

Appendix B: Replication of Previous Analyses for Posts from Five Days Before and After the Pledge Day

Table B1. Replication of the Main Analyses We Report in Sections 3.3.1 to 3.3.3

Analysis	T test			Chow test	
	P value	T value	DF	p value	F statistic
Sentiment	0.004	-2.654	155.47	0.798	0.227
Emotion-sadness	0.959	-0.051	153.72	0.346	1.062
Emotion-joy	0.046	2.012	153.31	0.0002	9.068
Emotion-fear	0.995	0.006	154.92	0.442	0.8168
Emotion-disgust	$< 10^{-9}$	-6.736	165.92	$< 10^{-12}$	28.44
Emotion-anger	0.181	1.345	149.86	0.790	0.235
Stock prices	$< 10^{-15}$	10.63	142.55	$< 10^{-15}$	363.9

Table B2. Replication of the Main Analysis We Report in Section 3.3.4

Random effects

Groups	Variance	Std. dev.
Date	1.2037	1.0971
Residual (ϵ)	0.1409	0.3753

Fixed effects

Coefficients	Estimate	Std. error	T value	P value	Confidence intervals	
					2.5%	97.5%
Intercept (β_0)	56.74	0.4918	115.385	$< 10^{-13}$	55.79	
SentimentScore (β_1)	-0.0148	0.0125	-1.188	0.235	-0.0392	
AnnouncementBefore (β_2)	-1.8138	0.6947	-2.611	0.031	-3.1586	

ICC (Date): 0.8952
Marginal R^2: 0.117
Conditional R^2: 0.907

About the Authors

Arthur Carvalho is the Dinesh & Ila Paliwal Innovation Chair and an Assistant Professor of Information Systems and Analytics at Farmer School of Business, Miami University. His research interests lie at the intersection of Information Systems, Analytics, and Decision Analysis. In particular, he works on the development and deployment of AI techniques and disruptive technologies, such as Hadoop services and Blockchain, aiming at improving business processes and decision making. His research has been published in academic journals such as Decision Support Systems, Decision Analysis, EURO Journal on Decision Processes, Judgement and Decision Making, The Journal of Prediction Markets, among others.

Adam Levitt is a graduate student in the Master of Accountancy program at Farmer School of Business, Miami University. Previously, he worked within the Treasury function at Progressive Insurance in a data analyst role. His areas of interest include data mining and accounting.

Seth Levitt is a graduate student in the Master of Accountancy program at Farmer School of Business, Miami University. Previously, he worked for Deloitte Advisory, focusing on assurance and internal audit engagements. His academic interests include big data and accounting.

Edward Khaddam is a graduate student in the Master of Science in Business Analytics at Haslam College of Business, University of Tennessee. His academic interests include big data and data mining.

John Benamati is a Professor of Information Systems and Chair of the Information Systems and Analytics Department at Farmer School of Business, Miami University. His major research and teaching interests are e-commerce trust, changing IT, information privacy, and IT management/strategy.