Routledge
Taylor & Francis Group

Journal of Management Information Systems

ISSN: 0742-1222 (Print) 1557-928X (Online) Journal homepage: http://www.tandfonline.com/loi/mmis20

Targeted Twitter Sentiment Analysis for Brands Using Supervised Feature Engineering and the Dynamic Architecture for Artificial Neural Networks

Manoochehr Ghiassi, David Zimbra & Sean Lee

Targeted Twitter Sentiment Analysis for Brands Using Supervised Feature Engineering and the Dynamic Architecture for Artificial Neural Networks

MANOOCHEHR GHIASSI, DAVID ZIMBRA, AND SEAN LEE

MANOOCHEHR GHIASSI (mghiassi@scu.edu) is a professor of information systems and Breetwor Fellow at Leavey School of Business, Santa Clara University. His research addresses machine learning, neural networks, artificial intelligence, business analytics, and software engineering. He has published widely in *IEEE, ACM, IIE, Computer Design*, and *Operations Research*.

DAVID ZIMBRA (dzimbra@scu.edu; corresponding author) is an assistant professor of OMIS at the Leavey School of Business at Santa Clara University. He received his Ph.D. from the University of Arizona. His research areas include social media analytics, sentiment analysis, machine learning, and business analytics.

SEAN LEE (sslee43@gmail.com) is a senior development engineer for Stella Technology. He received his M.S. in information systems from Santa Clara University. His research interests include information retrieval, opinion mining, and machine learning.

ABSTRACT: Social media communications offer valuable feedback to firms about their brands. We present a targeted approach to Twitter sentiment analysis for brands using supervised feature engineering and the dynamic architecture for artificial neural networks. The proposed approach addresses challenges associated with the unique characteristics of the Twitter language and brand-related tweet sentiment class distribution. We demonstrate its effectiveness on Twitter data sets related to two distinctive brands. The supervised feature engineering for brands offers final tweet feature representations of only seven dimensions with greater feature density. Reducing the dimensionality of the representations reduces the complexity of the classification problem and feature sparsity. Two sets of experiments are conducted for each brand in three-class and five-class tweet sentiment classification. We examine five-class classification to target the mild sentiment expressions that are of particular interest to firms and brand management practitioners. We compare the proposed approach to the performances of two state-of-the-art Twitter sentiment analysis systems from the academic and commercial domains. The results indicate that it outperforms these state-of-the-art systems by wide margins, with classification F_1-measures as high as 88 percent and excellent recall of tweets expressing mild sentiments. Furthermore, they demonstrate the tweet feature representations, though consisting of only seven dimensions, are highly effective in capturing indicators of Twitter sentiment expression. The proposed approach and vast majority of features identified through supervised

Journal of Management Information Systems / 2016, Vol. 33, No. 4, pp. 1034–1058.

ISSN 0742–1222 (print) / ISSN 1557–928X (online)
DOI: 10.1080/07421222.2016.1267526

feature engineering are applicable across brands, allowing researchers and brand management practitioners to quickly generate highly effective tweet feature representations for Twitter sentiment analysis on other brands.

Key words and phrases: artificial neural networks, feature engineering, sentiment analysis, social media, supervised feature engineering, Twitter.

Social media offer valuable information to firms about their consumers and brands. Consumers generate social media as a way to develop their relationships with brands [6]. A firm's brand-related social media activities can foster growth in their consumer base, increasing revenue [39]. The opinions (sentiments) expressed by consumers of brands in social media, conveyed in either text-based messages [27] or "likes" [23], can affect product sales. Researchers have advocated for the monitoring and management of these communications, and the development of automated approaches to social media analysis [7, 14, 24].

Twitter has emerged as a major social media platform, with more than 100 million users generating over 500 million tweets per day [37]. Tweets often express a user's perspective and opinion on a topic of interest, and research has shown they provide valuable insights on issues related to a firm's brand [18]. Accordingly, Twitter has generated great interest from sentiment analysis researchers, as well as firms and brand management practitioners. In spite of this attention and a growing body of literature, state-of-the-art Twitter sentiment analysis (TSA) approaches continue to perform poorly, with classification accuracies frequently below 70 percent [2]. Firms require accurate TSA to access the valuable consumer feedback communicated in social media, react to negative outcomes and publicity, and positively influence opinions of a brand.

These poor TSA performances may be attributed to several properties of tweets that make the TSA problem particularly challenging. Tweets are brief communications, limited to 140 characters in length, and characterized by diverse, evolving language with frequent use of slang, abbreviations, and emoticons. The brevity of tweets offers relatively few terms to evaluate with a sentiment lexicon, or resulting in sparsely populated tweet feature representations. Sparsity is a shortcoming of traditional feature representations and typically diminishes the performances of sentiment analysis methods.

The sentiment class distribution of the tweets of interest also complicates the analysis. Generally, the vast majority of tweets express neutral or no sentiment, and TSA approaches typically face the challenge of recalling the infrequent occurrences of positive or negative sentiment expressions. However, brand-related tweets frequently express a user's opinion of a brand, and these tend to be strong sentiments in one direction or the other [12]. From a brand management perspective, it is of particular interest to identify the subset of consumers whose perspective on the brand may be influenced and improved. Consumers expressing strongly positive sentiments are satisfied, and those expressing strongly negative sentiments may be entrenched in their position and resilient to marketing influence. However, consumers expressing mild sentiments may have opinions of a brand that could be improved. With these considerations, the tweets expressing mild sentiments toward a brand (positive or negative) should be targeted, but the majority of TSA approaches model the problem as a three-way classification (positive /negative /neutral). When brand-related tweet sentiment classes are further divided into strong and mild intensities, the tweets of interest (expressing mild sentiments) are typically far outnumbered by strong sentiments. Most machine-learning models lack the sensitivity to distinguish strong sentiment expressions from mild, and this strong class imbalance intensifies this deficiency. Without the ability to identify mild sentiment expressions, limited actionable intelligence can be provided to brand management practitioners from the TSA.

In this research, we present a targeted approach to TSA for brands that addresses the challenges associated with the unique characteristics of the Twitter language and the recall of mild sentiment expressions that are of particular interest to firms and brand management practitioners. Our methodology follows the design science paradigm of information systems research [16]. We hypothesize that among the reasons for the poor performances of state-of-the-art TSA approaches are the tweet feature representations used in their analyses. In our approach, we carefully craft tweet feature representations through supervised feature engineering for brands, resulting in final representations consisting of only seven dimensions with greater feature density. These feature representations are coupled with the dynamic architecture for artificial neural networks (DAN2), a machine-learning model with the sensitivity to distinguish mild sentiment expressions in tweet sentiment classification [12]. To isolate and evaluate the performances of DAN2, we also utilize the same tweet feature representations with another prominent machine-learning model, the support vector machine (SVM). In our experimentation with two distinctive brand-related Twitter data sets, we perform both three-class and five-class tweet sentiment classification (using multiple binary classifiers) to identify the mild expressions of sentiment. The results indicate the machine-learning models that use the tweet feature representations derived through the proposed supervised feature engineering approach for brands outperform state-of-the-art TSA systems from the academic (Sentiment140) [13, 34] and commercial (Repustate) [33] domains by wide margins, with classification precisions and recalls often above 80 percent. Furthermore, the DAN2 machine-learning model outperformed SVM in nearly every brand case and sentiment class, demonstrating particularly excellent recall of tweets expressing mild sentiments.

This study makes the following contributions to existing research. We develop a reduced feature representation specific to TSA. Reducing the dimensionality of the tweet feature representation also reduces problem complexity, increasing the density of the feature matrix, and mitigating the classical feature sparsity problem. We expand the number of sentiment classes from three to five, including mildly positive and mildly negative classes, to target the mild sentiment expressions of particular interest to firms and brand management practitioners. The study also offers DAN2 as a machine-learning model for tweet sentiment classification, a model with the sensitivity required to distinguish mild sentiment expressions and cope with the unbalanced class distribution typical of brand-related Twitter data sets and intensified by the five-sentiment class division. We evaluate the proposed approach on two brand-related cases in this study, and with the brand-related case examined in prior research [12], are developing a general and reusable feature set for TSA that can be applied across domains. Researchers simply need to augment this feature set with a small number of domain-specific features to generate a highly effective tweet feature representation for TSA in a new brand-related case. Our experimentation in this study and the prior study [12] has shown that 85 percent of the features used to generate the tweet feature representations are applicable in each brand-related case.

Related Research

Twitter sentiment analysis is a specialized problem within sentiment analysis, a prominent area of research in the field of computational linguistics. Approaches to sentiment analysis identify and evaluate opinions expressed in text using automated methods. Twitter has been the subject of much recent sentiment analysis research because tweets often express a user's perspective or opinion on an issue of interest, and the large volume of communications offers an unprecedented opportunity to derive valuable insights regarding firms and brands. There are several properties of tweets that make the TSA problem particularly challenging, including a diverse, informal, and evolving language and strong imbalance in the sentiment class distribution. Despite these unique challenges, the majority of approaches to TSA follow those developed for more traditional genres of communication like news articles [35], product reviews [7, 32], and web forums [1]. Traditional approaches to sentiment analysis can be broadly categorized into two classes. The first class of approaches involves the use of a lexicon of opinion-related terms in conjunction with a scoring method to evaluate the opinion expressed in text in an unsupervised application [20, 36]. These methods are widely applicable and fairly accurate, but their performance is limited as they are unable to account for contextual information, novel vocabulary, or other more nuanced indicators of opinion expression. The second class of approaches quantifies the text based on a feature representation, and applies a machine-learning algorithm to derive the relationship between the feature values and the opinion expressed through supervised learning [7, 32]. Models based on supervised learning require a large set of training instances complete with opinion class labels to calibrate model parameters. Finally, unnecessary, redundant, or infrequently occurring features in the feature representation introduce noise and diminish classification performance.

While many recent approaches to TSA continue to follow those developed for more traditional genres of communication, innovations designed to address the unique challenges associated with the problem have been proposed in the literature. Researchers have developed specialized tweet preprocessing procedures to remove or correct slang, abbreviations, misspellings, and exaggerations, and transform the Twitter language into a more traditional form [3, 29, 31]. Others have specifically leveraged these features common to the Twitter language, as well as emoticons, user mentions, hashtags, and hyperlinks, in the tweet feature representation [4, 5, 17, 22]. Another technique devised for TSA is to expand the number of available training instances by considering emoticons as noisy class labels [5, 13, 22, 31]. Researchers manually classify emoticons based on their interpreted sentiment expression, then collect and classify tweets containing the emoticons. The emoticon-based sentiment classifications are then used as noisy class labels to train a machine-learned classifier. This type of machine learning has been described as distant supervision [13]. In addition to expanding the number of available training instances, researchers have also improved performance by devising algorithms to expand the tweet feature representation. The brevity of tweets provides few terms to evaluate, and these algorithms are designed to generate additional potential indicators of sentiment in the expanded representation. Researchers have also used WordNet to supplement the content of tweets by leveraging the semantic relations of the words in the tweet, and including their synonyms, hypernyms, and antonyms [30]. The bootstrap parametric ensemble framework [15] identified an effective sentiment classifier by searching among the available Twitter data sets for training, features used to represent the tweet, and machine-learning classifiers.

This research builds on Ghiassi et al. [12], in which the authors introduced a supervised feature reduction approach using *n*-gram analysis, and defined a Twitter-specific feature representation for the sentiment analysis of tweets associated with a popular entertainer. The feature representation derived for the entertainer-brand-related tweets consisted of only 187 features but maintained good coverage over the Twitter data set (97 percent of tweets included at least one of the features). The derived tweet feature representation was coupled with the DAN2 machine-learned model for tweet sentiment classification and produced excellent performances, surpassing SVM, and alternative representations based on bag-of-words and the Opinion Finder lexicon [38].

Targeted Twitter Sentiment Analysis for Brands

In this research, we present a targeted approach to TSA for brands that addresses the challenges associated with the unique characteristics of the Twitter language and the recall of mild sentiment expressions that are of particular interest to firms and brand management practitioners. We extend the supervised feature engineering approach presented in Ghiassi et al. [12], and develop three additional stages designed to improve the tweet feature representations and model performances in brand-related TSA. The three new feature engineering stages are negation and valence shifter analysis, feature sentiment scoring, and aspect categorization. The supervised feature engineering for brands generates final tweet feature representations consisting of only seven dimensions with greater feature density. These feature representations are coupled with the dynamic architecture for artificial neural networks (DAN2), a machine-learning model with the sensitivity to distinguish mild sentiment expressions in tweet sentiment classification [12].

A preliminary demonstration of the proposed approach on a single brand was presented in Zimbra et al. [40]. In this research, we evaluate the proposed approach on two distinctive brands, related to a retail food provider and a prominent politician. We selected a well-known corporate brand that has been the focus of prior TSA research, Starbucks [18], and a very different brand, that of a politician generating much Twitter discussion with diverse opinions, Governor Chris Christie.

The conceptual design for our targeted TSA system for brands is presented in Figure 1. It consists of five major stages: tweet collection, tweet cleansing, feature engineering, feature vectorization, and tweet sentiment classification. We describe each of these stages in detail in the subsequent sections. For each brand-related case (Starbucks and Governor Christie), we collected a large number of tweets from the Twitter API, and randomly selected samples for our sentiment analysis. These data sets were preprocessed and cleansed. The cleansed tweets then underwent manual analysis and classification according to sentiment expression. These tweets complete with sentiment class labels were then used for the sentiment classification experimentation.

Brand-Related Twitter Data Collection and Cleansing

We collected tweets containing the @Starbucks or @GovChristie handles from the Twitter API to develop each brand-related tweet data set. During three months from August 18, 2013, to November 6, 2013, 442,443 tweets @Starbucks and 201,821 tweets @GovChristie were collected.

To cleanse the brand-related tweet data sets, we first removed all retweets, which contain previously communicated information. We next removed all tweets with multiple Twitter user handles, since the target of the sentiment expression may be unclear in these cases. Simple regular expression scripts were developed to identify and remove "RT" tweets and those with multiple "@usernames." For most brand-

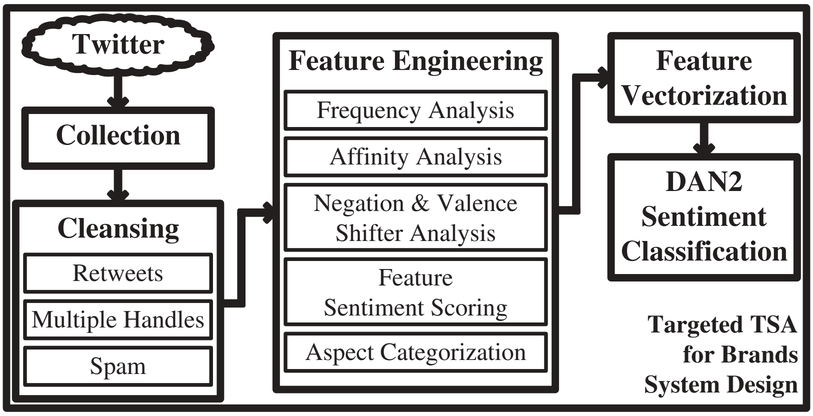

Figure 1. Targeted TSA for Brands System Design

related cases, retweets constitute the largest volume of tweets in a Twitter collection. Spam tweets were then manually identified and removed from the data sets. Spam removal is a more challenging data-cleansing task. In general, we considered spam to be tweets intended for promotional purposes. To identify spam tweets for removal, spam features for promotional tweets were defined during the manual scoring process. These spam features were then used to create regular expression scripts to remove other spam promotional tweets with the same feature. Some examples of spam features and promotional tweets from the @Starbucks data set are included below:

> Spam feature: "#RelaxWithPrincess" (157 instances)
> Example tweet: "follow @PrincessCruises by 5 pm PST to enter to win a $25 @Starbucks gift card! #RelaxWithPrincess"
> Spam feature: "win a" (2,078 instances)
> Example tweet: "Starbucks FLASH #Giveaway! Today only, enter to #win a $25 @Starbucks GC #MissionGiveaway http://t.co/kOyJgHvOWx"

Once the data sets were cleansed, random selections of tweets were drawn for further analysis and manual sentiment classification. Following these steps, the data sets for Starbucks and Governor Christie consisted of 9,367 and 4,302 tweets, respectively.

Each of these tweets was then manually analyzed and classified for sentiment expression by three graduate business students, into one of five sentiment classes: strongly positive, mildly positive, neutral, mildly negative, and strongly negative. Five sentiment classes were used to identify the mild expressions of sentiment that are of paramount interest to analysts and practitioners. Description of the sentiment classes and examples are presented in Table 1. Tweets that were unanimously classified by the three graduate business students were then used for the sentiment analysis experimentation, with the others causing disagreement removed from the data sets. Finally, 4,905 tweets remained in the Starbucks data set and 1,756 in the Governor Christie data set, complete with classifications into one of four sentiment classes. The remaining tweets, 4,462 for Starbucks and 2,546 for Governor Christie, were either neutral or did not receive unanimous sentiment classifications by the graduate business students, and were removed from the data sets. The sentiment class distribution of tweets in the two brand-related data sets are presented in Table 2.

Supervised Feature Engineering

The proposed approach to TSA for brands addresses the challenges associated with the unique characteristics of the Twitter language and the recall of mild sentiment expressions that are of particular interest to analysts and practitioners. We first describe the approach to developing the tweet feature representation through supervised feature engineering, then present DAN2 for sentiment analysis and classification.

Table 1. Description of the Tweet Sentiment Classes

Sentiment class	Description	Starbucks Tweets example	Governor Christie example
Strongly positive	Author clearly loves the subject	The last 2 seasons of my year are defined by @starbucks: Fall = 1st day of Pumpkin Spice Latte. Winter = 1st day of Christmas Blend. #truth	@GovChristie like I said we have the coolest Gov in America…#STTS it will be a blowout in Nov like never before and we in NJ can't wait
Mildly positive	Author likes the subject	Somehow we are staying at the only hotel in the vicinity that doesn't have a @Starbucks. This was very poor planning.	Ahh. @GovChristie, glad to see you understand #liberty and the #2A so well. I feel that this strain of #bigGovt is dangerous…
Neutral	Unclear how the author feels about the subject	So does someone know how to use the pitcher box of passion tea from @Starbucks? I'm really confused. Or I just can't read.	@GovChristie is headed to Asbury Park..
Mildly negative	Author dislikes the subject	This mornings #Crossfit workout was hard. But this coffee @Starbucks is harder. #AddSomeMoreH20Yo	@GovChristie Disabled vet booted from NJ boardwalk for bringing his service dog Nice how NJ treats Vets. #PutVets1st
Strongly negative	Author clearly hates the subject	@Starbucks I had a horrible experience at your store bad customer service the employee was very rude #Fail	@GovChristie you're just brown nosing your Dem supporters by signing this bill. Thanks 4 all your help n killing the US

Table 2 Sentiment Class Distribution for the Brand-related Twitter Data Sets

Sentiment class	Starbucks tweets	Governor Christie tweets
Strongly positive	2,861	642
Mildly positive	560	235
Mildly negative	798	386
Strongly negative	686	493
Total	4,905	1,756

The goal in supervised feature engineering is to select a set of features representing the tweet's content and the sentiment expressed. This approach produces a set of vocabulary, emoticons, and emojis for modeling and analysis. Similar to other modeling processes, a large set of features adds to the complexity of the classification problem. Adding to this challenge is the process of feature vectorization of the Twitter data set for input to a machine-learning model. The traditional feature engineering approach often fails to reduce the feature set, resulting in a sparse vectorization of the tweets. We address this specific challenge and present a supervised feature engineering approach in which we carefully craft the tweet feature representation to use in the TSA. Two overall objectives drive the decisions made in the feature engineering. The first objective is to include a sufficient number of features that will provide coverage over the tweet data set. In order for sentiment to be detected in a tweet, the tweet must contain one or more features included in the representation. Second, the feature engineering should reduce the number of dimensions in the feature space to reduce modeling complexity, create a denser representation, and alleviate sparsity. Superfluous parameters in a machine-learned model introduce noise and diminish sentiment classification performance.

The Twitter data sets used in this study are collected to be sufficient for the TSA models. Overall, we offer significantly reduced tweet feature representations and apply them to train multiple tweet sentiment classification models. The immediate advantage of a reduced, yet effective feature representation, is a reduction in the required size of the Twitter data set. And since the tweets are manually labeled, a smaller data set makes TSA more practical. However, we ensure that the Twitter data sets are sufficiently large to avoid overfitting. We emphasize that the final tweet feature representations include only seven "aspect" features. This significant dimensionality reduction provides TSA researchers the ability to study brand-related Twitter sentiments with manageably small data sets required for manual labeling. We also note that the feature set used to form the aspect features are mostly domain-agnostic. Only a few features (less than 15–20 percent) are domain-specific and particular to a brand. Similar to all artificial neural network (ANN)-based model validation processes, we retain separate testing datasets to measure the validity of our models. The results clearly indicate that the supervised feature engineering approach presented in this research is effective and accurate.

Our approach to supervised feature engineering builds upon the work of Ghiassi et al. [12], and consists of five stages: frequency analysis, affinity analysis, negation and valence shifter analysis, feature sentiment scoring, and aspect categorization. The first two stages were adapted from the feature engineering approach developed in prior research [12], while the last three stages are new advancements introduced in this study. We next discuss each of the supervised feature engineering stages in detail.

Supervised Feature Engineering: Frequency Analysis

The first stage in the feature engineering is frequency analysis, which focuses on the unigrams used in tweets and achieving the first overall objective of providing coverage over each tweet data set. All unigrams were extracted from the tweets in the data sets, and their frequencies of occurrences were counted. For instance, 52 percent of tweets in the Starbucks and 48 percent in Governor Christie data sets contained an emoticon or emoji. We evaluated the emoticons and emojis extracted and identified 47 emoticons and 131 emojis that definitively express a positive or negative sentiment. These 178 emoticons and emojis were included in both tweet feature representations. We then examined the frequencies of word unigrams, and removed infrequently occurring unigrams, while retaining a sufficient number to ensure 95 percent of the tweets in each data set contained at least one of the included word unigrams. This translated to a requirement of usage in about 0.01 percent of tweets in each data set, and resulted in approximately 1,300 and 1,200 word unigrams included in the tweet feature representation for Starbucks and Governor Christie, respectively. Upon further scrutiny, many of the words that were eliminated based on the frequency thresholds were synonyms of words included in the representations. To incorporate these synonymous terms and enrich the tweet feature representations, we manually develop synonym groups of word unigrams and apply the frequency thresholds at the group level instead of the word level. For example, awful, horrible, terrible, dreadful, atrocious, and horrendous composed a synonym group.

We examine emoticons and emojis separately and assess their specific sentiment polarities. We note that not all emoticons and emojis carry sentiment, they are used like icons. However, most emoticons and emojis do map to a positive or negative sentiment value (i.e., a heart emoji represents love, and a broken heart represents disappointment). We create two meta-features representing positive and negative emoticons and emojis. This feature reduction process is similar to that applied to the text. Finally, we use emoticons and emojis as features only and do not assume that their polarities represent the sentiment of the entire tweet.

Supervised Feature Engineering: Affinity Analysis

The second stage in the feature engineering is affinity analysis, which focuses on introducing higher order word *n*-grams into the tweet feature representations. Word *n*-grams contain rich sentiment expressions at the phrase level. To identify word phrases, we use the affinity measure [19], as defined below. For our data sets, word *n*-grams up to five words in length were considered in the affinity analysis.

$$Affinity(P) = f(P)/min_{\forall w_i \in P}(f(w_i)),$$

where $f(P)$ is the frequency of phrase P; $\min(f(w_i))$ is the minimum frequency across the words in phase P.

Affinity analysis identifies phrases containing words that frequently occur together in sequence. These phrases contain more complex and valuable sentiment expressions than their constituent word unigrams. To incorporate the word phrases while controlling the expansion of the tweet feature representation, we add the word phrases with high affinity but remove the constituent word unigrams from the representations. For example, from the Starbucks tweets presented in Table 1, the affinity analysis identified higher order word *n*-grams such as “Pumpkin Spice Latte,” “Christmas Blend,” “passion tea,” “horrible experience,” and “customer service”.

Supervised Feature Engineering: Negation and Valence Shifter Analysis

The third stage in the feature engineering is negation and valence shifter analysis, which focuses on adding negated, intensified, and diminished forms of the wordgram features identified in prior stages to the tweet feature representations. We manually examined the occurrences of negation of the wordgrams already included in the representations, and added any frequently occurring negated forms as additional features. The General Inquirer dictionary of overstatement and understatement words was then used as a reference to identify sentiment intensification or diminishment. The tweets in the data sets were scanned for occurrences of words used to intensify or diminish the wordgrams included in the representation. Frequently occurring intensified or diminished forms of the wordgrams were added to the tweet representation as additional features. For example, this stage identified intensified word *n*-grams like "very poor," "really confused," and "very rude" from the Starbucks tweets presented in Table 1.

Supervised Feature Engineering: Feature Sentiment Scoring

The fourth stage in the feature engineering is feature sentiment scoring, which focuses on assigning sentiment polarity and intensity scores to the terms included in the tweet feature representations. The tweet sentiment classifications manually assigned to the data sets were leveraged in the sentiment scoring. The information gain for each feature and tweet sentiment class was calculated, and based on these values features were assigned to one of seven sentiment groups: extremely positive, very positive, somewhat positive, neutral, somewhat negative, very negative, and extremely negative. When coupled with machine-learned classifiers, features from these groups were weighted with 16, 8, 4, 0, –4.1, –8.1, –20.1 intensities, respectively. The additional intensity weight and offset assigned to terms in negative sentiment groups is designed to ensure that negative tweets requiring the intervention of brand managers are not mistakenly classified. For instance, from the example Starbucks tweets presented in Table 1 the word *n*-grams "seasons" and "Pumpkin Spice Latte" were assigned to the extremely positive sentiment group, while "horrible experience" and "very rude" were assigned to the extremely negative sentiment group.

Sarcasm is a challenge in any sentiment analysis. The automatic identification and evaluation of sarcasm is very difficult [25, 28]. Literature discussing sarcasm in TSA suggests using "keywords, or hash tagged words" to analyze the presence of sarcasm. We have followed this approach and have selected a number key words (such as "thanks," "#smh," #not") that are followed by negative (or positive) sentiment markers, and assigned a slightly higher (or lower) weight to the negative (or positive) feature to counterweigh the presence of the sarcastic qualifying term. These rules generate some automation in the process, and along with manual examination of the tweets, we manage sarcasm effectively.

Supervised Feature Engineering: Aspect Categorization

The fifth and final stage in the feature engineering is aspect categorization. Sentiment analysis literature defines subjects as "entities" with attributes or "aspects" [26]. In this context, sentiment can be applied to an entity or to attributes or aspects of the entity. Liu [26] extends the definition of aspect by introducing aspect categories and aspect expressions. In related work, Kontopoulos et al. [21] developed an ontology-based approach to the sentiment analysis of consumer comments on smartphones. They offer an ontology and state that "the domain of reference is semantically divided and sentiment scores are assigned not to the whole statement (i.e., tweets), but to various aspects of the topic at hand" [21, p. 4067].

In this research, we follow that principle but differ in the interpretation and implementation of the aspect concept. We offer a set of seven aspects that describe the various ways users express sentiments regarding the brands. These aspect categories are desire, interjections, quality, review, transactional, domain-specific, and ancillary. We map each of the tweet features to an aspect category, collapsing the final representations to only seven dimensions. We do not simply assign a sentiment score to each aspect, but compute a score as the sum of the sentiment scores of all features belonging to that aspect. This sum reflects the collective sentiment of features belonging to this aspect. The assignment of polarity scores to features, and mapping of features to aspects allows for the automation of this step. In our definition an aspect can be thought of as a "meta feature" representing a class of features referring to a similar brand-related aspect. The premise is that a feature's effect on a tweet's sentiment is accounted for through its belonging to the aspect, and the semantic relationship that aspect holds with the sentiment class of the tweet. Therefore, the information content of an individual feature and its sentiment is carried through the "meta feature," as represented by the seven aspects. The major benefits of this approach are dimensionality reduction (reducing many features into seven aspects), significant reduction of sparsity in the feature matrix, and reduction of noise introduced to the machine-learning model. For example from the Starbucks tweets presented in Table 1, the word grams "very poor" and "hard" are mapped to the quality aspect category, "customer service" is mapped to the transactional aspect category, and "Pumpkin Spice Latte," "Christmas Blend," and "passion tea" are mapped to the domain-specific aspect category. Finally, four aspects were defined generally so they could be applied for TSA across many subjects (Twitter handles) or brands, regardless of the domain of analysis: desire, review, transactional, and quality aspects. The remaining aspects were created for domain specificity, to handle negations, or to capture specific emotions (the interjection aspect).

Sentiment Classification: Dynamic Architecture for Artificial Neural Networks

Following supervised feature engineering, the tweet feature representation values are provided as input to the DAN2 [10] for sentiment analysis and classification, a machine-learning model with the sensitivity to distinguish mild sentiment expressions in tweets that are of interest to firms and brand management practitioners. DAN2 has been successfully applied in prior studies to TSA [12], text classification [8, 9], and time series forecasting [11]. We next present an overview of DAN2.

DAN2 employs an architecture different from the traditional neural network (FFBP) models. The general philosophy of the DAN2 model is based on the principle of learning and accumulating knowledge at each layer, propagating and adjusting this knowledge forward to the next layer, and repeating these steps until the desired network performance criteria are reached. Figure 2 presents the overall DAN2 architecture. As in classical neural networks, the DAN2 architecture is composed of an input layer, hidden layers, and an output layer. The input layer accepts external data to the model. In DAN2, unlike classical neural nets, the number of hidden layers is not fixed a priori. They are sequentially and dynamically generated until a level of performance accuracy is reached. In addition, the proposed approach uses a fixed number of hidden nodes (four) in each hidden layer. This structure is not arbitrary, but justified by the estimation approach. At each hidden layer, the network is trained using all observations in the training set simultaneously, so as to minimize a stated training accuracy measure such as mean squared error (MSE) value or other accuracy measures. As shown in Figure 2, each hidden layer is composed of four nodes. The first node is the bias or constant (e.g., 1) input node, referred to as the C node. The second node is a function that encapsulates the "current accumulated knowledge element" (CAKE node) during the previous training step. The third and fourth nodes represent the current residual (remaining) nonlinear component of the process via a transfer function of a weighted and

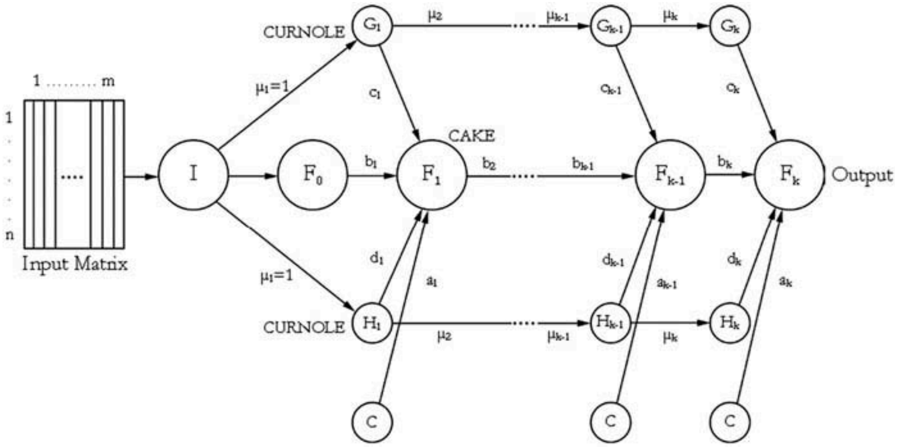

Figure 2. The DAN2 Network Architecture

normalized sum of the input variables. Such nodes represent the "current residual nonlinear element" (CURNOLE nodes). In Figure 2, the I node represents the input, the C nodes are the constant nodes, the G_k and H_k nodes represent CURNOLE nodes, and the F_k nodes are the CAKE nodes. The final CAKE node represents the dependent variable or the output. At each layer, the previous four nodes (C, G_k, H_k, *and* F_{k-1}) are used as the input to produce the next output value (F_k). The parameters on the arcs leading to the output nodes, (a_k, b_k, c_k, d_k), represent the weights of each input in the computation of the output for the next layer. The parameter connecting the CURNOLE nodes, μ_k, is used as part of the argument for the CURNOLE nodes and reflects the relative contribution of each input vector to the final output values at each layer. A detailed description of the architecture and its properties are presented in [10]. The training process begins with a special layer where the CAKE node captures the linear component of the input data. Thus, its input (content) is a linear combination (weighted sum) of the input variables and a constant input node. These weights are easily obtainable through classical linear regression. If the desired level of accuracy is reached, we can conclude that the relationship is linear and the training process stops. This step is used as the starting point. For classification problems this step can be replaced with an alternative method. For nonlinear relations additional hidden layers are required. At each subsequent layer the input to the CAKE node is a weighted sum (linear combination) of the previous layer's CAKE, CURNOLE, and C nodes. Throughout training, the CAKE nodes carry an adequate portion of learning achieved in previous layers forward. This process ensures that the performance or knowledge gained so far is adjusted and improved but not lost. This property of DAN2 introduces knowledge memorization to the model. Ghiassi and Saidane [10] show that the DAN2 algorithm ensures that during network training, the residual error is reduced in every iteration and the accumulated knowledge is monotonically increased.

The training process defines creation of partitions among classes that could include linear and nonlinear components. The linear component of the input data is captured in the first CAKE node using ordinary least squares (OLS) or other simple and easy to compute approaches. The algorithm next transforms the input data set to model the nonlinearity of the process in subsequent iterations. DAN2 uses a vector projection approach to perform data transformation. The transformation process defines a reference vector $R = \{r_j;\ j = 1, 2, \ldots, m\}$, where m represents the number of attributes of the observation records, and projects each observation record onto this vector to normalize the data, as discussed in Ghiassi and Saidane [10]. This normalization defines an angle, α_i, between records i and the reference vector R. DAN2 uses the set of α_i's to train the network, and updates their values in every iteration. Ghiassi and Saidane [10] show that this normalization can be represented by the trigonometric function Cosine ($\mu_k\ \alpha_i + \theta_k$). In every hidden layer k of the architecture we vary ($\mu_k\ \alpha_i + \theta_k$) and measure the impact of this change on the output value. The modification of the angle ($\mu_k\ \alpha_i + \theta_k$) is equivalent to rotating μ_k and shifting θ_k the reference vector, thus changing the impact of the projected input vectors and their contribution to the output for that iteration. The Cosine ($\mu_k\ \alpha_i + \theta_k$)

uses two (nonlinear) parameters, μ_k and θ_k. The use of the latter can be avoided through the expansion of the cosine function in the form: A Cosine (μ_k α_i) + B Sine (μ_k α_i). We use this functional form as the transfer function in our model. The two CURNOLE nodes in Figure 2 represent this formulation. At any given hidden layer *k*, if the Cosine (μ_k α_i + θ_k) terms captured in previous layers do not adequately express the nonlinear behavior of the process, a new layer with an additional set of nodes is automatically generated, including a new Cosine (μ_k α_i + θ_k) term. This process is analogous to how the Fourier series adds new terms to improve function approximation. Therefore, the number of layers in the DAN2 architecture is dynamically defined and depends on the complexity of the underlying process and the desired level of accuracy. Thus, the output of this model is represented by the linear combination of the constant, CAKE, and CURNOLE nodes. Equation (1) represents the functional form of this relationship at iteration (layer) *k*:

$$F_k(X_i) = a_k + b_k F_{k-1}(X_i) + c_k G_k(X_i) + d_k H_k(X_i) \tag{1}$$

where X_i represents the *n* independent input records, $F_k(X_i)$ represents the output value at layer *k*, $G_k(X_i)$ = Cosine (μ_k α_i), and $H_k(\mu_k\ \alpha_i)$ = Sine (μ_k α_i) represent the transferred nonlinear components, and a_k, b_k, c_k, d_k, and μ_k are parameter values at iteration *k*. The training process initially captures the linear component by using ordinary least squares (OLS) or other simple and easy to compute approaches. If the desired level of accuracy is reached, the training terminates. Otherwise, the model generates additional layers to capture the nonlinear component of the process by minimizing a measure of total error as represented by $SSE_k = \Sigma_i[F_k(X_i) - F\hat{}(X_i)]^2$. Substituting $F_k(X_i)$ from Equation (1) results in:

$$SSE_k = \sum\nolimits_i [a_k + b_k F_{k-1}(X_i) + c_k \mathrm{Cos}(\mu_k \alpha_i) + d_k \mathrm{Sin}(\mu_k \alpha_i) - F\hat{}(X_i)]^2 \tag{2}$$

where $F\hat{}(X_i)$ are the observed output values. Minimizing Equation (2) requires the estimation of five parameters. This formulation is linear in the parameter set A_k, where $A_k = \{a_k, b_k, c_k, d_k\}$ and nonlinear in parameter μ_k. Ghiassi and Saidane [10] present several nonlinear optimization strategies to estimate the nonlinear parameter μ_k. We also show that following this approach, at each layer the knowledge gained is monotonically increased, total error is reduced, and the network training improves.

DAN2's algorithm is highly scalable. The algorithm is composed of a series of layers that are automatically and dynamically generated. At each layer the observation vectors are projected into the reference vector to create the angle α_i as discussed earlier. The training and optimization at each layer uses only the transformed data to estimate the five parameters of Equation (2). Therefore, regardless of the original size of the data set, DAN2 only needs to compute the set of four linear parameters, $A_k = \{a_k, b_k, c_k, d_k\}$, and one nonlinear parameter μ_k at each iteration. For instance, for text classification problems with large feature sets, once the starting point is computed, each observation record, as represented by its many features, is projected onto the reference vector and its angle, α_i, is computed. The training process uses only the α_i values and needs to compute only the five parameters. Therefore, the original problem with a large feature set is converted to a five-parameter model at each layer; thus scaling the problem size and demonstrating the scalability of the model. The scalability of DAN2 is a distinguishing strength of the approach from traditional artificial neural networks. Ghiassi and Saidane [10] compare DAN2 with traditional feed-forward back-propagation (FFBP) and recurrent neural network (RNN) models. The comparison spans both theoretical and computational perspectives using several benchmark data sets from the literature. Performance of DAN2 against these models as well as nonneural network alternatives is also presented. Their studies show that DAN2 outperforms all other alternatives and produces more accurate training and testing results in every case.

Twitter Sentiment Analysis Experimentation

To evaluate the effectiveness of the proposed approach to TSA, two sets of experiments were conducted for each brand. Since the vast majority of TSA approaches provide three-class classifications (positive /negative /neutral), we first perform experiments in simple positive and negative sentiment classification. We then perform five-class sentiment classification experiments, to evaluate the ability to identify the mild expressions of sentiment that are of paramount interest to analysts and practitioners.

In the experimentation, we compare the proposed approach to the performances of two state-of-the-art TSA systems, Sentiment140 from the academic domain [13, 34] and a commercial system, Repustate [33]. The Sentiment140 system [34] uses a maximum entropy-based machine-learned classifier trained on a large Twitter corpus using distant supervision, a technique in which emoticons are used as noisy sentiment class labels for tweets in the training. Word and part-of-speech *n*-grams are used as the features to represent tweets. Sentiment140 outputs three-class (positive /negative /neutral) tweet sentiment classifications. Limited information was available on the sentiment analysis approach applied by the Repustate system [33] because it is a proprietary commercial offering. The Repustate system was selected for the experimentation due to its prominence in the sentiment analysis market, and because it outputs continuous sentiment scores rather than discrete three-class classifications. To conduct experiments in three-class and five-class sentiment classification these continuous scores were mapped to sentiment classes. Training data were used to determine the class boundary thresholds to apply to the sentiment scores. We ranked the sentiment scores output by the Repustate system, and identified the thresholds in score that would maintain the sentiment class distribution observed in the training data set. For example, if there were 100 strongly positive instances in the training data set, the top 100 sentiment scores for the training data were considered to be associated with this class, and the lowest score among these would serve as the threshold dividing the strongly positive class from the weakly positive class, and so on for the remaining sentiment classes. The sentiment score class boundary thresholds established using the training data set were then applied in evaluation on the testing data.

To demonstrate the effectiveness of DAN2 for sentiment analysis and classification, we also develop comparable SVM models that are provided the very same features and instances as input. For DAN2 and SVM, multiple binary classifiers were developed to perform one versus all sentiment classification and identify instances belonging to a specific sentiment class. For three-class sentiment classification, two binary classifiers were developed to identify positive or negative instances. And for five-class classification, binary classifiers were developed to identify strongly positive, weakly positive, weakly negative, and strongly negative instances. Neutral sentiment classifiers were not developed; if an instance was not positively classified by one of the sentiment classifiers, it was considered to be neutral.

Since the brand-related data sets were strongly unbalanced in sentiment class distribution, training and testing data sets were carefully developed before use in the calibration and evaluation of TSA approaches. We first randomly split the two data sets into training and testing sets of about 70–80 percent and 30–20 percent, respectively. Since strongly positive sentiment classes far outnumber the other sentiment classes in each data set, all available out-of-class instances were used in the development of the strongly positive sentiment classifiers. For the other sentiment classes in each brand, training and testing data sets were scaled so the number of in-class instances represented at least 30 percent of the total number of instances, to provide sufficient in-class exposure to the machine-learned models. Out-of-class instances were randomly selected, but required to maintain sentiment class distributions representative of each data set overall.

The training and testing data sets for the Starbucks and Governor Christie TSA experimentation are presented in Table 3 for three-class classification, and Table 4 for five-class classification. As an example, Table 5 presents the distribution of tweets used in the training and testing of the mildly negative sentiment classifiers for Starbucks. Table 6 presents the distribution of tweets used in the strongly positive sentiment classifiers for Governor Christie.

Twitter Sentiment Analysis Results

The results of experimentation in three-class and five-class tweet sentiment classification are presented in Tables 7–12 below. The models labeled DAN2 and SVM used the tweet feature representation derived through the feature engineering

Table 3. Training and Testing Data for Three-Class Sentiment Classification

	Starbucks		Governor Christie	
Sentiment classifier	Training tweets	Testing tweets	Training tweets	Testing tweets
Positive	2,734	687	702	175
Negative	1,144	340	703	176
Total	3,878	1,027	1,405	351

Table 4. Training and Testing Data for Five-Class Sentiment Classification

	Starbucks		Governor Christie	
Sentiment classifier	Training tweets	Testing tweets	Training tweets	Testing tweets
Strongly positive	2,281	580	514	128
Mildly positive	453	107	188	47
Mildly negative	605	193	309	77
Strongly negative	539	147	394	99
Total	3,878	1,027	1,405	351

Table 5. Tweet Distribution for the Mildly Negative Sentiment Classifiers for Starbucks

Tweets	Training	Testing	Total
Strongly positive tweets	694	169	863
Mildly positive tweets	134	37	171
Mildly negative tweets	622	154	776
Strongly negative tweets	172	42	214
Total	1,622	402	2,024

Table 6. Tweet Distribution for the Strongly Positive Sentiment Classifiers for Christie

Tweets	Training	Testing	Total
Strongly positive tweets	514	128	642
Mildly positive tweets	188	47	235
Mildly negative tweets	309	77	386
Strongly negative tweets	394	99	493
Total	1,405	351	1,756

described in this research, coupled with either the DAN2 or SVM machine-learned models for sentiment analysis and classification, respectively. Also listed are the results generated by the Sentiment140 and Repustate TSA systems. Precision, recall, and F_1-measure are reported for three-class and five-class tweet sentiment classification. Since Sentiment140 performs only three-class sentiment classification, no results are provided for the five-class classification.

The precision, recall, and F_1-measure statistics for the three-class tweet sentiment classification are presented in Tables 7, 8, and 9, respectively. As the results show, the tweet feature representations derived through the proposed approach to supervised feature engineering for brands were highly effective in capturing indicators of sentiment expression, though consisting of only seven dimensions. The DAN2 and

Table 7. Class-Level Precision for Three-Class Sentiment Classification

Brand	Sentiment class	DAN2, %	SVM, %	Sentiment140, %	Repustate, %
Starbucks	Positive	84.2	66.2	75.7	73.6
	Negative	85.6	75.9	40.0	45.3
Governor Christie	Positive	82.0	60.3	82.0	71.9
	Negative	88.1	78.7	85.5	78.7

Table 8. Class-Level Recall for Three-Class Sentiment Classification

Brand	Sentiment class	DAN2, %	SVM, %	Sentiment140, %	Repustate, %
Starbucks	Positive	83.0	69.1	39.6	46.5
	Negative	86.0	56.8	24.7	45.8
Governor Christie	Positive	83.9	59.3	51.6	68.6
	Negative	81.4	83.7	20.2	51.8

Table 9. Class-Level F_1-Measure for Three-Class Sentiment Classification

Brand	Sentiment class	DAN2, %	SVM, %	Sentiment140, %	Repustate, %
Starbucks	Positive	83.5	67.6	52.0	57.0
	Negative	85.7	64.7	30.5	45.6
Governor Christie	Positive	82.7	59.8	63.3	70.2
	Negative	84.4	80.6	32.6	62.5

Table 10. Class-Level Precision for Five-Class Sentiment Classification

Brand	Sentiment class	DAN2, %	SVM, %	Repustate, %
Starbucks	Strongly positive	85.9	84.6	67.3
	Mildly positive	77.9	0.9	12.6
	Mildly negative	79.3	58.1	19.7
	Strongly negative	91.9	93.8	33.9
Governor Christie	Strongly positive	82.1	84.2	62.8
	Mildly positive	81.6	0.0	17.6
	Mildly negative	81.3	67.6	29.1
	Strongly negative	94.7	89.3	49.7

SVM models that used these representations performed well, and outperformed Sentiment140 and Repustate.

Overall, the Sentiment140 and Repustate TSA systems performed poorly. The systems demonstrated fair precision in tweet sentiment classification, with over 71 percent precision for each brand and sentiment class (except for the Starbucks

Table 11. Class-Level Recall for Five-Class Sentiment Classification

Brand	Sentiment class	DAN2, %	SVM, %	Repustate, %
Starbucks	Strongly positive	87.6	88.3	36.8
	Mildly positive	66.7	0.9	13.2
	Mildly negative	87.0	37.0	20.0
	Strongly negative	85.1	76.7	34.2
Governor Christie	Strongly positive	89.8	82.8	59.9
	Mildly positive	68.9	0.0	16.8
	Mildly negative	85.1	86.2	19.1
	Strongly negative	78.0	81.3	32.8

Table 12. Class-Level F_1-Measure for Five-Class Sentiment Classification

Brand	Sentiment class	DAN2, %	SVM, %	Repustate, %
Starbucks	Strongly positive	86.8	86.4	47.6
	Mildly positive	71.8	0.9	12.9
	Mildly negative	83.0	45.2	19.9
	Strongly negative	88.4	84.4	34.1
Governor Christie	Strongly positive	85.8	83.5	61.3
	Mildly positive	74.7	0.0	17.2
	Mildly negative	83.2	75.8	23.1
	Strongly negative	85.5	85.1	39.5

negative class). However, the recall of these systems was poor, ranging from 20 percent to 68 percent, demonstrating difficulties in identifying the diverse sentiment expressions encountered in TSA. These poor performances in recall lowered their F_1-measures, which ranged from 30 percent to 70 percent. Comparatively, the SVM models that used tweet feature representations derived through the supervised feature engineering for brands outperformed Sentiment140 and Repustate in F_1-measure for each brand and sentiment class (except for the Governor Christie positive class). SVM model F_1-measures ranged from 59 percent to 80 percent.

Furthermore, the DAN2 model performances were superior to SVM, while using the same tweet feature representations, producing better results in terms of precision, recall, and F_1-measure for each brand and sentiment class (except for negative class recall of Governor Christie). DAN2 model F_1-measures ranged from 82 percent to 85 percent, demonstrating excellent consistency in performances across brand cases and sentiment classes. DAN2 also far outperformed Sentiment140 and Repustate for each brand and sentiment class. The F_1-measures of DAN2 models in three-class tweet sentiment classification were each over 82 percent, improving upon the Sentiment140 and Repustate systems by more than 12 percent for each brand and sentiment class. The DAN2 improvements over SVM were similarly pronounced; an increase in F_1-measure of over 15 percent for each brand and sentiment class with the exception of the Governor Christie negative class, in which the DAN2 improvement over SVM was about 4 percent. For each brand and sentiment class, the DAN2 models produced the best three-class tweet sentiment classification performances and F_1-measures among all competing models and systems.

To assess the statistical significance of these improvements in tweet sentiment classification recall produced by the proposed approach, pairwise t-tests were conducted on the classification experimentation results for both the Starbucks and Governor Christie brand cases (alpha = 0.05; two-tailed tests). In three-class tweet sentiment classification for each brand, the models that used the supervised feature engineering for brands and DAN2 produced highly significant improvements in recall of each sentiment class, compared to the Sentiment140 and Repustate systems ($p <$ 0.001). DAN2 also produced significant improvements over SVM models that applied the same tweet feature representations for each brand and sentiment class ($p < 0.001$; except for the Governor Christie negative class).

The precision, recall, and F_1-measure statistics for the five-class tweet sentiment classification are presented in Tables 10, 11, and 12, respectively. Five-class tweet sentiment classification targets the mild sentiment expressions of particular interest to firms and brand management practitioners. Further demonstrated by these five-class results, the tweet feature representations derived through supervised feature engineering for brands were highly effective in capturing indicators of sentiment expression. The DAN2 and SVM models that used these representations again outperformed the state-of-the-art commercial TSA system Repustate.

Again in five-class classification, the Repustate system performed poorly overall. The system demonstrated better precision than recall, but experienced difficulties in identifying the diverse sentiment expressions encountered in TSA. Repustate system precisions ranged from 12 percent to 67 percent, while recalls ranged from 13 percent to 59 percent. Classifying mild sentiment expressions was particularly challenging, with lower precisions ranged from 12 percent to 29 percent, and recalls from 13 percent to 20 percent. The Repustate system performed better in classifying the strong sentiment classes.

The SVM models that used tweet feature representations derived through supervised feature engineering for brands outperformed Repustate terms of precision, recall, and F_1-measure for each brand and sentiment class (except mildly positive classes). However, the SVM models experienced similar difficulties in classifying mild sentiment expressions. For mild sentiment classes, SVM model precisions ranged from 0 percent to 67 percent and recalls ranged from 0 percent to 86 percent; for strong sentiment classes, precisions ranged from 84 percent to 93 percent and recalls ranged from 76 percent to 88 percent.

Consistent with the three-class results, DAN2 produced the best model performances in five-class tweet sentiment classification overall, superior to SVM models in F_1-measures for each brand and sentiment class while using the same tweet feature representations. DAN2 models also outperformed Repustate by all measures for each brand and sentiment class. The more complex five-class tweet sentiment classification revealed the sensitivity required from machine-learning models to distinguish the infrequently occurring mild sentiment expressions that are of particular interest to firms and brand management practitioners. The superiority of DAN2 over SVM in this application is further supported in the mild sentiments classification experimentation results. In terms of precision for mild sentiment classes, DAN2 performances ranged from 77 percent to 81 percent, improvements over SVM models ranging from 13 percent to 81 percent and over the Repustate system ranging from 52 percent to 65 percent. DAN2 model performances in recall ranged from 66 percent to 87 percent for mild sentiment classes, substantial improvements over SVM models in all but the mildly negative sentiment class for Governor Christie, and over the Repustate system ranging from 52 percent to 66 percent. Finally in terms of F_1-measure for mild sentiment classes, DAN2 performances ranged from 71 percent to 83 percent, improvements over SVM models ranging from 7 percent to 74 percent and over the Repustate system ranging from 57 percent to 63 percent. For each brand and sentiment class, the DAN2 models produced the best five-class tweet sentiment classification performances and F_1-measures among the competing approaches.

We similarly assessed the statistical significance of the improvements in sentiment class-level recall produced by the proposed approach by conducting pairwise t-tests for both the Starbucks and Governor Christie brand cases. In five-class tweet sentiment classification for each brand, the models that utilized the supervised feature engineering for brands and DAN2 produced highly significant improvements in recall compared to the Sentiment140 and Repustate systems ($p < 0.001$). DAN2 also produced significant improvements over SVM models that applied the same tweet feature representations in five of the eight brand sentiment classes ($p < 0.001$).

Discussion and Conclusions

In this research, we presented a targeted approach to TSA for brands using supervised feature engineering and DAN2. The approach addresses the challenges associated with the unique characteristics of the Twitter language and the recall of mild sentiment expressions that are of particular interest to firms and brand management practitioners. We demonstrated the effectiveness of the proposed approach on two Twitter data sets related to the Starbucks and Governor Chris Christie brands. We crafted tweet feature representations through supervised feature engineering, resulting in final representations consisting of only seven dimensions with greater feature density. These feature representations were coupled with DAN2, a machine-learning model with the sensitivity to distinguish mild sentiment expressions in tweet sentiment classification. We compared the proposed approach to the performances of several competitors in three-class and five-class tweet sentiment classification, including two state-of-the-art TSA systems from the academic (Sentiment140) [13, 34] and commercial (Repustate) [33] domains. To isolate and evaluate the performances of DAN2, we also used the same tweet feature representations with another prominent machine-learning model, the support vector machine (SVM).

The results indicated the DAN2 and SVM machine-learning models that used the tweet feature representations derived through the proposed supervised feature engineering approach for brands outperformed the state-of-the-art Sentiment140 and Repustate TSA systems by wide margins, with classification precisions and recalls often above 80 percent. Furthermore, the DAN2 machine-learning model outperformed SVM in nearly every brand case and sentiment class, demonstrating particularly excellent recall of tweets expressing mild sentiments. In both three- and five-class tweet sentiment classification, the improvements in tweet classification recall provided by the proposed approach that used the engineered tweet feature representations and DAN2 over Sentiment140 and Repustate were statistically significant for each brand case and sentiment class. DAN2 also produced significant improvements over SVM models that applied the same tweet feature representations for most brand cases and sentiment classes. The DAN2 models produced the best performances and F_1-measures in three- and five-class tweet sentiment classification among all competing models and systems for each brand and sentiment class, with F_1-measures often above 80 percent.

This study made the following contributions to existing research. We developed a reduced feature representation specific to TSA. Reducing the dimensionality of the tweet feature representation also reduced problem complexity, increased the density of the feature matrix, and mitigated the classical feature sparsity problem. We expanded the number of sentiment classes from three to five, included mildly positive and mildly negative classes, to target the mild sentiment expressions of particular interest to firms and brand management practitioners. We also further evaluated DAN2 as a machine-learning model for tweet sentiment classification. DAN2 demonstrated the sensitivity required to distinguish mild sentiment expressions and cope with the unbalanced class distribution typical of brand-related Twitter data sets and intensified by the five-sentiment class division. We evaluated the proposed approach on two brand-related cases in this study, and with the brand-related case examined in prior research [12], continued to develop a general and reusable feature set for TSA that can be applied across domains. Researchers simply need to augment this feature set with a small number of domain-specific features to generate a highly effective tweet feature representation for TSA in a new brand-related case. Our experimentation revealed that 85 percent of the features used to generate the tweet feature representations are applicable in each brand-related case.

There are several limitations and future directions to our research. While numbering in the thousands and of sufficient sizes to train a machine-learned classifier, our Twitter data sets are limited in size. Manual annotation of tweets to create gold-standard sentiment class labels for classifier training and evaluation is a laborious and time-consuming task, and this limits the size of our brand-related Twitter data sets. We will continue to explore additional brand-related cases and Twitter data sets in future research. Our approach to supervised feature engineering will also continue to improve in future research. A challenge to the approach described in this study (and others in the state of the art) is dealing with sarcasm in sentiment expression. We will continue to advance our supervised feature engineering approach to develop improved tweet feature representations and performances in TSA.

References

1. Abbasi, A.; Chen, H.; and Salem, A. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems*, *26*, 3 (2008), 1–34.

2. Abbasi, A.; Hassan, A.; and Dhar, M. Benchmarking Twitter sentiment analysis tools. In *Proceedings of the Language Resources and Evaluation Conference*, Reykjavik, Iceland: LREC, 2014, pp. 823–829.

3. Agarwal, A.; Xie, B.; Vovsha, I.; Rambow, O.; and Passonneau, R. Sentiment analysis of Twitter data. In *Proceedings of the ACL Human Language Technologies Conference*, 2011, pp. 30–38.

4. Bakliwal, A.; Foster, J.; van der Puil, J.; O'Brien, R.; Tounsi, L.; and Hughes, M. Sentiment analysis of political tweets: Towards an accurate classifier. In *Proceedings of the ACL Workshop on Language in Social Media*, Atlanta, GA: LASM, 2013, pp. 49–58.

5. Barbosa, L., and Feng, J. Robust sentiment detection on Twitter from biased and noisy data. In *Proceedings of the International Conference on Computational Linguistics*, Beijing, China: COLING, 2010, pp. 36–44.

6. Chen, A.; Lu, Y.; Chau, P.; and Gupta, S. Classifying, measuring, and predicting users' overall active behavior on social networking sites. *Journal of Management Information Systems*, *31*, 3 (2014), 213–253.

7. Gamon, M. Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the Conference on Computational Linguistics*, Geneva, Switzerland: COLING, 2004, p. 841.

8. Ghiassi, M., and Burnley, C. Measuring effectiveness of a dynamic artificial neural network algorithm for classification problems. *Expert Systems with Applications*, *37*, 4 (2010), 3118–3128.

9. Ghiassi, M.; Olschimke, M.; Moon, B.; and Arnaudo, P. Automated text classification using a dynamic artificial neural network model. *Expert Systems with Applications*, *39*, 12 (2012), 10967–10976.

10. Ghiassi, M., and Saidane, H. A dynamic architecture for artificial neural network. *Neurocomputing, 63* (2005), 397–413.

11. Ghiassi, M.; Saidane, H.; and Zimbra, D. K. A dynamic artificial neural network model for forecasting time series events. *International Journal of Forecasting*, *21*, 2 (2005), 341–362.

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

13. Go, A.; Bhayani, R.; and Huang, L. Twitter sentiment classification using distant supervision. Technical Report, Stanford Digital Library Technologies Project, 2009.

14. Godes, D.; Mayzlin, D.; Chen, Y.; Das, S.; Dellarocas, C.; Pfeiffer, B.; Libai, B.; Sen, S.; Shi, M.; and Verlegh, P. The firm's management of social interactions. *Marketing Letters*, *16*, 3–4 (2005), 415–428.

15. Hassan, A.; Abbasi, A.; and Zeng, D. Twitter sentiment analysis: A bootstrap ensemble framework. In *Proceedings of the ASE/IEEE Conference on Social Computing*, Washington, D.C.: ASE/IEEE, 2013, pp. 357–364.

16. Hevner, A.; March, S.; Park, J.; and Ram, S. Design science in information systems research. *MIS Quarterly*, *28*, 1 (2004), 75–105.

17. Hu, Y.; Wang, F.; and Kambhampati, S. Listening to the crowd: Automated analysis of events via aggregated Twitter sentiment. In *Proceedings of the Conference on Artificial Intelligence*, Las Vegas, NV, 2013, pp. 2640–2646.

18. Jansen, B.; Zhang, M.; Sobel, K.; and Chowdury, A. Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, *60*, 11 (2009), 2169–2188.

19. Kajanan, S.; Shariff, A.; Datta, A.; Dutta, K.; and Paul, D. Twitter post filter for mobile applications. In *Proceedings of the Workshop on Information Technology and Systems*, Shanghai, China, 2011.

20. Kim, S., and Hovy, E. Determining the sentiment of opinions. In *Proceedings of the International Conference on Computational Linguistics*, Geneva, Switzerland: COLING, 2004, p. 1367.

21. Kontopoulos, E.; Berberidis, C.; Dergiades, T.; and Bassiliades, N. Ontology-based sentiment analysis of Twitter posts. *Expert Systems with Applications*, *40*, 10 (2013), 4065–4074.

22. Kouloumpis, E.; Wilson, T.; and Moore, J. Twitter sentiment analysis: The good the bad and the OMG! In *Proceedings of the AAAI Conference on Weblogs and Social Media*, Barcelona, Spain: AAAI, 2011, pp. 538–541.

23. Lee, K.; Lee, B.; and Oh, W. Thumbs up, sales up? The contingent effect of Facebook likes on sales performance in social commerce. *Journal of Management Information Systems*, *32*, 4 (2015), 109–143.

24. Li, X.; Sun, S.; Chen, K.; Fung, T.; and Wang, H. Design theory for market surveillance systems. *Journal of Management Information Systems*, *32*, 2 (2015), 278–313.

25. Liebrecht, C.; Kunneman, F.; and van den Bosch, A. The perfect solution for detecting sarcasm in tweets #not. In *Proceedings of the Conference on Computational Approaches to Subjectivity, Sentiment, and Social Media*, Atlanta, GA: ACL, 2013, pp. 29–37.

26. Liu, B. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. New York, NY: Cambridge University Press, 2015.

27. Liu, Y. Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing*, *70*, 3 (2006), 74–89.

28. Maynard, D., and Greenwood, M. Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. In *Proceedings of the Language Resources and Evaluation Conference*, 2014, pp. 4238–4243.

29. Mittal, A., and Goel, A. Stock prediction using Twitter sentiment analysis. *Working Paper, Stanford University*, 2012.

30. Montejo-Ráez, A.; Martínez-Cámara, E.; Martín-Valdivia, M. T.; and Ureña-López, L. A. Ranked WordNet graph for sentiment polarity classification in Twitter. *Computer Speech and Language*, *28*, 1 (2014), 93–107.

31. Pak, A., and Paroubek, P. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Language Resources and Evaluation Conference*, Valletta, Malta: LREC, 2010, pp. 1320–1326.

32. Pang, B.; Lee, L.; and Vaithyanathan, S. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing*, Strousburg, PA: ACL, 2002, pp. 79–86.

33. Repustate (2015). *www.repustate.com*.

34. Sentiment140 (2015). *www.sentiment140.com*.

35. Tetlock, P. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62, 3 (2007), 1139–1168.

36. Turney, P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the Association for Computational Linguistics Conference*, Philadelphia, PA: ACL, 2002, pp. 417–424.

37. Twitter, Inc. IPO prospectus. Downloaded on February 2, 2014, (http://www.sec.gov/Archives/edgar/data/1418091/000119312513390321/d564001ds1.htm), 2013.

38. Wilson, T.; Hoffman, P.; Somasundaran, S.; Kessler, J.; Wiebe, J.; Choi, Y.; Cardie, C.; Riloff, E.; and Patwardhan, S. OpinionFinder: A system for subjectivity analysis. In *Proceedings of the Conference on Human Language Technology - Empirical Methods in Natural Language Processing*, Vancouver, Canada: HLT/EMNLP, 2005, pp. 34–35.

39. Xie, L., and Lee, Y.J. Social media and brand purchase: Quantifying the effects of exposures to earned and owned social media activities in a two-stage decision-making model. *Journal of Management Information Systems*, *32*, 2 (2015), 204–238.

40. Zimbra, D.; Ghiassi, M.; and Lee, S. Brand-related Twitter sentiment analysis using feature engineering and the dynamic architecture for artificial neural networks. *Proceedings of the Hawaii International Conference on System Sciences*, *49*, Koloa, HI: IEEE, 2016, pp. 1930–1938.