2016 49th Hawaii International Conference on System Sciences

Brand-Related Twitter Sentiment Analysis using Feature Engineering and the Dynamic Architecture for Artificial Neural Networks

David Zimbra
Santa Clara University
dzimbra@scu.edu

M. Ghiassi
Santa Clara University
mghiassi@scu.edu

Sean Lee
Santa Clara University
sslee2@scu.edu

Abstract

We present an approach to brand-related Twitter sentiment analysis using feature engineering and the Dynamic Architecture for Artificial Neural Networks (DAN2). The approach addresses challenges associated with the unique characteristics of the Twitter language, and the recall of mild sentiment expressions that are of interest to brand management practitioners. We demonstrate the effectiveness of the approach on a Starbucks brand-related Twitter data set. The feature engineering produced a final tweet feature representation consisting of only seven dimensions, with greater feature density. Two sets of experiments were conducted in three-class and five-class tweet sentiment classification. 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 the approach outperforms these state-of-the-art systems in both three-class and five-class tweet sentiment classification by wide margins, with classification accuracies above 80% and excellent recall of mild sentiment tweets.

1. Introduction

Twitter has emerged as a major social media platform, with more than 100 million users generating over 500 million tweets per day [1]. 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 [2]. Accordingly, Twitter has generated great interest from sentiment analysis researchers, as well as brand management practitioners. In spite of this attention and a growing body of literature, state-of-the-art Twitter sentiment analysis approaches continue to perform poorly, with classification accuracies frequently below 70% [3].

These poor performances may be attributed to several properties of tweets that make the Twitter sentiment analysis 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 yields sparsely populated tweet feature representations. These conditions typically diminish 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 Twitter sentiment analysis approaches 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 [4]. From a brand management perspective, it is of paramount interest to identify the subset of consumers whose perspective on the brand may be influenced and improved. With these considerations, tweets expressing mild sentiments toward a brand should be targeted by Twitter sentiment analysis approaches. Consumers expressing strongly positive sentiments require no further intervention, and those expressing strongly negative sentiments may be entrenched in their position and more resilient to marketing influence. In spite of these practical considerations, the majority of Twitter sentiment analysis approaches model the problem as a three-way (positive/negative/neutral) classification. Furthermore, the state-of-the-art methods are often insufficiently sensitive to distinguish strong sentiment expressions from the mild, providing limited actionable intelligence to brand management practitioners.

In this research, we present an approach to brand-related Twitter sentiment analysis that addresses the challenges associated with the unique characteristics of the Twitter language and the recall of mild sentiment expressions. We believe that the poor performances of state-of-the-art Twitter sentiment analysis approaches may be attributed to the features considered in the analysis. In our approach we carefully craft the tweet

1530-1605/16 $31.00 © 2016 IEEE
DOI 10.1109/HICSS.2016.244

feature representation through supervised feature engineering, resulting in a final representation consisting of only seven dimensions. This tweet feature representation is coupled with the Dynamic Architecture for Artificial Neural Networks (DAN2) [5] for sentiment analysis and classification, a machine-learned model with sufficient sensitivity to distinguish mild sentiment expressions in tweets. In our experimentation with a Starbucks brand-related Twitter data set, we perform both three-class and five-class sentiment classification (using multiple binary classifiers) to identify the mild expressions of sentiment that are of particular interest to brand management practitioners. The results indicate that the proposed approach outperforms state-of-the-art Twitter sentiment analysis systems from the academic (Sentiment140) [6; 7] and commercial (Repustate) [8] domains by wide margins, with classification accuracies above 80% and excellent recall of mild sentiment tweets.

2. 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 as 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 business and society. There are several properties of tweets that make the Twitter sentiment analysis 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 Twitter sentiment analysis follow those developed for more traditional genres of communication like news articles [9], product reviews [10; 11], and web forums [12]. 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 [13; 14]. 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 quantify the text based upon a feature representation, and apply a machine learning algorithm to derive the relationship between the feature values and the opinion expressed through supervised learning [10; 11]. Models based upon supervised learning require a large set of training instances complete with opinion class labels to calibrate model parameters, and suffer from domain specificity restricting their potential for application. Also, unnecessary, redundant, or infrequently occurring features in the feature representation introduces noise and diminishes classification performance.

While many recent approaches to Twitter sentiment analysis 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 [15; 16; 17]. 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 [18; 19; 20; 21]. Another technique devised for Twitter sentiment analysis is to expand the number of available training instances by considering emoticons as noisy class labels [7; 15; 18; 19]. Researchers manually classify emoticons based upon 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 [7]. 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. Montejo-Ráez et al [22] utilized 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. The bootstrap parametric ensemble framework [23] 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 upon the work of Ghiassi et al [4], in which they 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% of tweets included at least one of the features). The derived tweet feature representation was coupled with the DAN2 machine-learned model for sentiment analysis and classification and produced excellent performance, far surpassing the performances of SVM, and alternative representations including bag-of-words and the Opinion Finder lexicon [24].

In this research we extend the feature engineering approach of Ghiassi et al [4], and develop three additional stages designed to further reduce the feature representation while maintaining good coverage over the Twitter data set. We evaluate the proposed approach on a newly-collected Twitter data set associated with a popular product-related brand. The three new feature engineering stages devised in this research are presented in Section 4, including negation and valence shifter analysis, feature sentiment scoring, and aspect categorization.

3. Twitter data

To evaluate the proposed approach to brand-related Twitter sentiment analysis, we selected a well-known brand that has been the focus of prior research in the area [2], Starbucks. We collected tweets containing the '@Starbucks' handle from the Twitter API for three months from August - October 2013, resulting in 442,443 tweets. Retweets were removed, which contain previously tweeted information, as well as tweets where multiple user handles were used, since the target of the sentiment expression may be unclear. 254,196 tweets remained. A random selection of 9,367 tweets was then drawn to be carefully analyzed and manually classified for sentiment by a team of three graduate students. Five sentiment classes were used in the classification, described by the scale in Table 1, in particular to identify the mild expressions of sentiment that are of paramount interest to brand management practitioners. Tweets that were unanimously assigned to a given sentiment class by all three evaluators were retained for the experiments. The final Starbucks data set consisted of 5,526 tweets. The sentiment class distribution for the data set is presented in Table 2.

Table 1. Description of tweet sentiment classes

Sentiment Class	Description	Example
Strongly Positive	Author clearly loves Starbucks	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
Mildly Positive	Author likes Starbucks	Somehow we are staying at the only hotel in the vicinity that doesn't have a @Starbucks. This was very poor planning.
Neutral	Unclear how the author feels about Starbucks	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.
Mildly Negative	Author dislikes Starbucks	This mornings #Crossfit workout was hard. But this coffee @Starbucks is harder. #AddSomeMoreH20Yo
Strongly Negative	Author clearly hates Starbucks	@Starbucks I had a horrible experience at your store bad customer service the employee was very rude #Fail

Table 2. Sentiment class distribution for Starbucks data set

Sentiment Class	Tweets
Strongly Positive	2885
Mildly Positive	617
Neutral	414
Mildly Negative	783
Strongly Negative	827
Total	5526

4. Twitter sentiment analysis

The proposed approach to brand-related Twitter sentiment analysis 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 brand management practitioners. We first describe the approach to developing the tweet feature representation through supervised feature engineering, then present the DAN2 model for sentiment analysis and classification.

Supervised feature engineering allows us to carefully craft the tweet feature representation to apply in the sentiment analysis. There are two overall objectives driving 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 a feature included in the representation. Second, the feature engineering should reduce the number of dimensions in the feature space to create a denser representation and alleviate sparsity.

Superfluous parameters in a machine-learned model introduces noise and diminishes sentiment classification performance. Our approach to supervised feature engineering builds upon the work of Ghiassi et al [4], 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 [4], while the last three stages are new advancements introduced in this research. We next discuss each of the feature engineering stages in detail.

4.1. 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 the Twitter data set. All unigrams were extracted from the tweets in the Starbucks data set, and their frequencies of occurrences counted. We began by examining emoticons and emoji, which can serve as fairly explicit indicators of sentiment. 52% of tweets in the Starbucks data set contain an emoticon or emoji. We evaluated the emoticons and emoji extracted and identified 47 emoticons and 131 emoji that definitively express a positive or negative sentiment. These 178 emoticons and emoji were included in the tweet feature representation. We then examined the frequencies of word unigrams, and removed infrequently occurring unigrams while retaining those required to ensure 95% of the tweets in the Starbucks data set contained at least one included word unigram. This translates to a frequency requirement of usage in 0.01% of tweets, and resulted in approximately 1,300 word unigrams included in the tweet feature representation. Upon further scrutiny, many of the words that were eliminated based upon the frequency threshold were synonyms of words included in the representation. To incorporate these synonymous terms and enrich the tweet feature representation, we manually developed synonym groups of word unigrams and apply the frequency threshold at the group level instead of the word level. For example, awful, horrible, terrible, dreadful, atrocious, and horrendous comprised a synonym group.

4.2. 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 representation. Word n-grams contain rich sentiment expressions at the phrase-level. To identify word phrases, we utilize the affinity measure [25], as defined below. For the Starbucks brand case, word n-grams up to five words in length were considered in the 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 representation. For example, from the tweets presented in Table 1 the affinity analysis stage identified higher order word n-grams like 'Pumpkin Spice Latte', 'Christmas Blend', 'passion tea', 'horrible experience', and 'customer service'.

4.3. 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 word gram features identified in prior stages to the tweet feature representation. We manually examined the occurrences of negation of the word grams already included in the representation, and added any frequently occurring negated forms as additional features. The General Inquirer dictionary of overstatement and understatement words was then used as reference to identify sentiment intensification or diminishment. The tweets in the Starbucks data set were scanned for occurrences of these words used to intensify or diminish the word grams already included in the representation. Frequently occurring intensified or diminished forms of the word grams 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 tweets presented in Table 1.

4.4. 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 representation. The tweet sentiment classifications manually assigned to the Starbucks data set were leveraged in the sentiment scoring. The Information Gain for each feature and tweet sentiment class was calculated, and based upon 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 example, from the 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.

4.5. Feature engineering: aspect categorization

The fifth and final stage in the feature engineering is aspect categorization, which focuses on reducing the dimensions of the tweet feature representation using a brand aspect ontology. Consumers often express sentiments toward a brand in terms of distinct brand-related aspects. In a related approach, Kontopoulos et al [26] developed an ontology-based approach to the sentiment analysis of consumer comments on smart phones. Evaluating sentiment in terms of brand-related aspects also provides more actionable intelligence to brand managers regarding consumer opinion. Based upon our scrutiny of the Starbucks data set, seven aspect categories were defined, describing the various ways that users express sentiments regarding the Starbucks brand. These aspect categories were desire, interjections, quality, review, transactional, domain-specific, and ancillary. We map each of the features in the tweet feature representation to an aspect category, collapsing the final representation to seven dimensions. Sentiment scores were summed across features within an aspect category, and these values were then provided as input to the machine-learned sentiment classifiers. The aspect categorization yields a representation with greater feature density. For example, from the tweets presented in Table 1 the word n-grams 'very poor' and 'hard' were mapped to the quality aspect category, 'customer service' was mapped to the transactional aspect category, and 'Pumpkin Spice Latte', 'Christmas Blend', and 'passion tea' were mapped to the domain-specific aspect category.

4.6. Sentiment analysis: dynamic architecture for artificial neural networks

Following feature engineering, the tweet feature representation values are provided as input to the Dynamic Architecture for Artificial Neural Networks (DAN2) [5] for sentiment analysis and classification, a machine-learned model with sufficient sensitivity to distinguish mild sentiment expressions in tweets of particular interest to brand management practitioners. DAN2 has been successfully applied in prior studies to Twitter sentiment analysis [4], text classification [27], and time series forecasting [28].

DAN2 employs a different architecture than the traditional artificial neural network (FFBP) models. The general philosophy of the DAN2 model is based upon 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. 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. Unlike classical neural networks, in DAN2 the number of hidden layers is not fixed a priori. They are sequentially and dynamically generated until a level of performance accuracy is reached. Additionally, DAN2 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). In Ghiassi and Saidane [5], the authors compare DAN2 with traditional FFBP and recurrent neural network (RNN) models, spanning theoretical, computational, and performance perspectives using several benchmark datasets from the literature. The results revealed DAN2 outperformed all comparative approaches and produced more accurate results in every case.

5. Sentiment analysis experimentation

To evaluate the proposed approach to brand-related Twitter sentiment analysis two sets of experiments were conducted. Since the vast majority of Twitter sentiment analysis 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 brand management practitioners.

In the experimentation, we compare the proposed approach to the performances of two state-of-the-art Twitter sentiment analysis systems, Sentiment140 from the academic domain [6; 7] and the commercial system Repustate [8]. The Sentiment140 system [7] uses a maximum entropy-based machine learned classifier trained on a large Twitter corpus using distant supervision, a technique where emoticons are used as noisy sentiment class labels for tweets in the training. Word and part-of-speech n-grams are utilized 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, as 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 was 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. These 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 Support Vector Machine (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 vs. 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 to identify strongly positive, weakly positive, weakly negative, and strongly negative instances were developed. 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 Starbucks data set was strongly unbalanced in its sentiment class distribution, training and testing data sets were carefully developed before use in the calibration and evaluation of Twitter sentiment analysis approaches. We first randomly split the data set into training and testing sets of about 80% and 20% respectively. Positive and strongly positive sentiment classes far outnumber the other sentiment classes, so all available out-of-class instances were used in the development of their sentiment classifiers. For the other sentiment classes, training and testing data sets were scaled so the number of in-class instances represented at least 30% of the total number of instances in the data set, to provide sufficient in-class exposure to machine-learned models. Out-of-class instances were randomly selected, but required to maintain a sentiment class distribution similar to the overall Starbucks data set. The training and testing data sets for three-class and five-class sentiment classification are presented in Table 3 and Table 4.

Table 3. Training and testing data for three-class sentiment classification

Sentiment Classifier	Training Tweets	Testing Tweets
Positive	4420	1106
Negative	4293	1073

Table 4. Training and testing data for five-class sentiment classification

Sentiment Classifier	Training Tweets	Testing Tweets
Strongly Positive	4420	1106
Mildly Positive	1645	411
Mildly Negative	2088	522
Strongly Negative	2205	551

6. Sentiment analysis results

The results of experimentation in three-class, and five-class tweet sentiment classification on the Starbucks data set are presented in Tables 5-7. The models labeled DAN2 and SVM utilized the tweet feature representation derived through the feature engineering described previously, 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 Twitter sentiment analysis systems.

In Table 5, the overall sentiment classification accuracies for the four approaches are presented. Since

Sentiment140 performs three-class sentiment classification, no results for the five-class problem are provided. As shown by the classification accuracies, the tweet feature representation derived through feature engineering was highly effective in capturing indicators of sentiment expression in the Starbucks data set despite consisting of seven dimensions. DAN2 and SVM models that utilized this feature representation outperformed the Sentiment140 and Repustate systems, with classification accuracies exceeding 78%. The DAN2 model performed the best overall, with 86% sentiment classification accuracy in the three-class problem, and 85% in five-class classification. The Sentiment140 and Repustate systems performed poorly overall, with classification accuracies around 40%.

To assess the statistical significance of the improvements in classification accuracy produced by the proposed feature engineering approach and DAN2, pair-wise t-tests were conducted on the experiment results (*n*=2179 for three-class, *n*=2590 for five-class; alpha=0.05; two-tailed tests). In both three-class and five-class experiments, the approaches that utilized feature engineering (DAN2 and SVM) produced highly significant improvements in accuracy (at $p<0.001$) compared to the Sentiment140 and Repustate systems. DAN2 models also significantly outperformed models that used SVM in classification accuracy in both three-class and five-class experiments (at $p<0.01$), while applying the exact same tweet feature representations.

Table 5. Overall sentiment classification accuracies

Sentiment Classes	DAN2	SVM	Sentiment140	Repustate
Three	86.06%	78.33%	39.96%	45.82%
Five	85.56%	78.39%	--	34.09%

The sentiment class-level recall statistics are also presented in Table 6 and Table 7 for the three-class, and five-class tweet sentiment classification experiments, respectively. Overall, the performances in sentiment class-level recall further demonstrate the effectiveness of the tweet feature representation derived through feature engineering, and DAN2 for sentiment analysis and classification. In the three-class experiment results shown in Table 6, the DAN2 and SVM sentiment class recall performances were far better and more consistent across positive and negative classes, compared with relatively poor recall from both the Repustate and Sentiment140 systems (around 40%), and inconsistency between sentiment class recalls from Sentiment140. The best performance in the three-class problem was DAN2, with 82% positive and 86% negative class recall.

Table 6. Class-level recall for three-class sentiment classification

Sentiment Class	DAN2	SVM	Sentiment140	Repustate
Positive	82.99%	79.26%	40.39%	46.78%
Negative	86.04%	78.35%	22.94%	40.74%

The recall performances in five-class tweet sentiment classification experiments are also presented in Table 7. While the DAN2 and SVM models that utilize the tweet feature representation again drastically outperformed the Repustate system, the effectiveness of DAN2 in distinguishing the mild sentiment expressions of interest to brand management practitioners becomes apparent. As shown by the mild sentiment class-level recalls, applying the tweet feature representation to DAN2 for sentiment analysis and classification resulted in improvements of 5% and 22% over SVM in mildly positive and mildly negative class recalls respectively. The DAN2 models were relatively consistent in recall across sentiment classes with performances in the upper 80%'s, with the exception of the mildly positive class. The SVM model and Repustate system had similar difficulties in recalling the mildly positive sentiment class.

Table 7. Class-level recall for five-class sentiment classification

Sentiment Class	DAN2	SVM	Repustate
Strongly Positive	87.59%	84.30%	40.20%
Mildly Positive	66.67%	61.40%	9.02%
Mildly Negative	87.01%	65.60%	21.93%
Strongly Negative	85.07%	91.10%	17.65%

We similarly assessed the statistical significance of the improvements in sentiment class-level recall produced by the proposed feature engineering approach and DAN2 by conducting pair-wise t-tests. In the three-class experiments, approaches that utilized feature engineering (DAN2 and SVM) produced significant improvements in positive and negative class recall (at $p<0.001$) compared to the Sentiment140 and Repustate systems. DAN2 models also outperformed models that used SVM in recall for both sentiment classes (at $p<0.01$). Significant improvements were also observed in the five-class experiments. The approaches that utilized feature engineering produced significant improvements in sentiment class recall for each of the four sentiment classes (at $p<0.001$) compared to the Repustate system. And with the exception of the strongly negative class, DAN2 models significantly outperformed models that used SVM in the recall of all other sentiment classes (at $p<0.01$), while applying the same tweet feature representations.

7. Discussion

In this research, we presented an approach to brand-related Twitter sentiment analysis using feature engineering and the Dynamic Architecture for Artificial Neural Networks (DAN2). The approach addressed challenges associated with the unique characteristics of the Twitter language and the recall of mild sentiment expressions that are of paramount interest to brand management practitioners. We demonstrated the effectiveness of the approach on a Starbucks brand-related Twitter data set consisting of 5,526 tweets. The feature engineering produced a final tweet feature representation consisting of only seven dimensions, with greater feature density. To evaluate the proposed approach, two sets of experiments were conducted in three-class and five-class tweet sentiment classification. In the experimentation, we compared the proposed approach to the performances of two state-of-the-art Twitter sentiment analysis systems from the academic and commercial domains. The results indicated that the proposed approach outperformed these state-of-the-art systems in both three-class and five-class tweet sentiment classification by wide margins, with classification accuracies above 80% and excellent recall of mild sentiment tweets (as well as strong sentiments).

Several conclusions emerged from these results. While only seven dimensions, the tweet feature representation effectively captured the expressions of sentiments in Starbucks brand-related tweets. The DAN2 and SVM models performed very well, and outperformed the state-of-the-art systems, using the representation derived through feature engineering. Furthermore, the performances of the DAN2 models were superior to the SVM models across all experiments and sentiment classes, in terms of classification accuracy and recall, with only one exception (strongly negative class recall). Importantly, DAN2 models demonstrated the sensitivity required to accurately distinguish both mild and strong sentiment expressions. Future research along this line will examine additional brands of great diversity.

8. References

[1] Twitter, Inc. (2013). IPO Prospectus. Downloaded on February 2, 2014, (*http://www.sec.gov/Archives/edgar/data/1418091/000119312513390321/d564001ds1.htm*).

[2] Jansen, B., Zhang, M., Sobel, K., and Chowdury, A. (2009). Twitter Power: Tweets as Electronic Word of Mouth. *Journal of the American Society for Information Science and Technology*, 60.

[3] Abbasi, A., Hassan, A., and Dhar, M. (2014). Benchmarking Twitter Sentiment Analysis Tools. *Proc of LREC Conf.*

[4] Ghiassi, M., Skinner, J., and Zimbra, D. (2013). Twitter Brand Sentiment Analysis: A Hybrid System using N-gram Analysis and Dynamic Artificial Neural Network. *Expert Systems with Applications*, 40.

[5] Ghiassi, M. and Saidane, H. (2005). A Dynamic Architecture for Artificial Neural Network. *Neurocomputing*, 63.

[6] Sentiment140 (2015). *www.sentiment140.com*.

[7] Go, A., Bhayani, R., and Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. *Technical Report, Stanford Digital Library Technologies Project*.

[8] Repustate (2015). *www.repustate.com*.

[9] Tetlock, P. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, 62.

[10] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs Up?: Sentiment Classification using Machine Learning Techniques. *Proc of the ACL Conf on EMNLP*.

[11] Gamon, M. (2004). Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis. *Proc of Conf on Computational Linguistics*.

[12] Abbasi, A., Chen, H., and Salem, A. (2008). Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. *ACM Transactions on Information Systems*, 26(3).

[13] Turney, P. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proc of ACL Conf.*

[14] Kim, S. and Hovy, E. (2004). Determining the Sentiment of Opinions. *Proc of the Intl Conf on Computational Linguistics*.

[15] Pak, A. and Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Proc of LREC Conf.*

[16] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. (2011). Sentiment Analysis of Twitter Data. *Proc of ACL HLT Conf.*

[17] Mittal, A. and Goel, A. (2012). Stock Prediction Using Twitter Sentiment Analysis. *Working Paper, Stanford University*.

[18] Barbosa, L. and Feng, J. (2010). Robust Sentiment Detection on Twitter from Biased and Noisy Data. *Proc of COLING Conf.*

[19] Kouloumpis, E., Wilson, T., and Moore, J. (2011) Twitter Sentiment Analysis: The Good the Bad and the OMG! Proc of AAAI Conf on Weblogs and Social Media.

[20] Bakliwal, A., Foster, J., van der Puil, J., O'Brien, R., Tounsi, L., and Hughes, M. (2013). Sentiment Analysis of Political Tweets: Towards an Accurate Classifier. *Proc of ACL Workshop on Language in Social Media*.

[21] Hu, Y., Wang, F., and Kambhampati, S. (2013). Listening to the Crowd: Automated Analysis of Events via Aggregated Twitter Sentiment. *Proc of Conf on Artificial Intelligence*.

[22] Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M. T., and Ureña-López, L. A. (2014). Ranked WordNet Graph for Sentiment Polarity Classification in Twitter. *Computer Speech and Language*, 28.

[23] Hassan, A., Abbasi, A., and Zeng, D. (2013). Twitter Sentiment Analysis: A Bootstrap Ensemble Framework, *Proc of the ASE/IEEE Conf on Social Computing*.

[24] Wilson, T., Hoffman, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. (2005). OpinionFinder: A System for Subjectivity Analysis. Proc of Conf on HLT-EMNLP.

[25] Kajanan, S., Shafeeq Bin Mohd Shariff, A., Datta, A., Dutta, K., and Paul, D. (2011). Twitter Post Filter for Mobile Applications. *Proc of the Workshop on Information Technology and Systems*.

[26] Kontopoulos, E., Berberidis, C., Dergiades, T., and Bassiliades, N. (2013). Ontology-Based Sentiment Analysis of Twitter Posts. *Expert Systems with Applications*.

[27] Ghiassi, M., Olschimke, M., Moon, B., and Arnaudo, P. (2012). Automated Text Classification using a Dynamic Artificial Neural Network Model. *Expert Systems with Applications*, 39.

[28] Ghiassi, M., Saidane, H., and Zimbra, D. K. (2005). A Dynamic Artificial Neural Network Model for Forecasting Time Series Events. *International Journal of Forecasting*, 21.