Predicting Customer Satisfaction in Customer Support Conversations in Social Media Using Affective Features

Jonathan Herzig, Guy Feigenblat,
Michal Shmueli-Scheuer,
David Konopnicki
IBM Research - Haifa
Haifa 31905, Israel
{hjon,guyf,shmueli,davidko}@il.ibm.com

Anat Rafaeli
Technion-Israel Institute of Technology
Haifa 32000, Israel
anatr@ie.technion.ac.il

ABSTRACT

Providing customer support through social media channels is gaining popularity. In such a context, predicting customer satisfaction in an early stage of a service conversation is important. Such an analysis can help personalize agent assignment to maximize customer satisfaction, and prioritize conversations. In this paper, we show that affective features such as customer's and agent's personality traits and emotion expression improve prediction of customer satisfaction when added to more typical text based features. We only utilize information extracted from the first customer conversation turn and previous customer and agent social network activity. Thus, our customer satisfaction classifier outputs its prediction in an early stage of the conversation, before any interaction has taken place between the customer and an agent. Our model was trained and tested on a Twitter conversations dataset of two customer support services, and shows an improvement of 30% in F1-score for predicting dissatisfaction.

Keywords

Affective computing; classification; customer support

1. INTRODUCTION

As part of the raging societal and commercial success of social media, applications go far beyond the initial use case of person to person communication. Social media are rapidly becoming an integral part of corporate Customer Relationship Management. In this context, an interesting use case for social media is customer support, which used to assume a private conversation between a customer and a service rep (agent), and can now take place over public social media channels. A recent study shows that one in five customers in the U.S (23%) say they have used social media for customer support in 2014, up from 17% in 2012[1]. Obviously, companies hope that such uses are associated with a positive experience. Yet, there are limited tools for assessing this.

In this work we explore the relation between *affect* evident in a conversation and satisfaction with customer support provided through social media. Our objective is to predict customer satisfaction given affective evaluations of both customers and agents. Specifically we model *affect* by considering *personality traits* and *emotions*, two aspects of individuals' intrinsic dispositions that are different in temporality: personality traits are relatively long term and permanent while emotions are relatively short term and transient. In the context of customer support, it was shown that customers tend to express negative emotions such as frustration and disappointment, as well as positive emotions such as gratitude [5]. As to *personality traits*, many studies examined the effect of specific traits of agents with respect to customer satisfaction, as well as how to interpret traits of customers. For example, customer trust and compromising traits correlate with positive satisfaction [14, 3].

With the advance of behavioral studies in social media, online services are available for assessing the *personality traits* of social media users based on their online interactions (e.g., tweets, forum posts)[2,3]. These services use text analytics to infer personality and social characteristics. As to *emotions*, analysis services based on textual messages are gaining popularity both in academic studies (cf. [16, 15]), and industry[4,5] as a method to get valuable insights about a person from their textual content.

Using these capabilities, our goal is to predict customer satisfaction from the initiation of the interaction (i.e., the first message that is posted by a customer), given the customer's and the selected agent's *personality traits* (obtained from their social media history prior to the conversation), and the *emotions* expressed by the customer in this first message. These data are utilized to predict the satisfaction of the customer from the entire conversation. Companies that provide customer support can foremost benefit from such a prediction and use it, for example in their agent assignment process (to assign an agent with personality traits

UMAP '16, July 13-17, 2016, Halifax, NS, Canada
© 2016 ACM. ISBN 978-1-4503-4370-1/16/07...$15.00
DOI: http://dx.doi.org/10.1145/2930238.2930285

[1] http://about.americanexpress.com/news/docs/2014x/2014-Global-Customer-Service-Barometer-US.pdf
[2] http://analyzewords.com
[3] https://watson-pi-demo.mybluemix.net/
[4] http://www.sentimetrix.com/
[5] http://apidemo.theysay.io/

that will maximize customer satisfaction), or to provide the assigned agent information about the state of the customer (e.g., the customer is angry, and not open to changes). To our knowledge, this is the first research that shows how to utilize *affect* of both parties of a conversation in order to increase the customer support satisfaction provided through social media.

2. RELATED WORK

Various works have studied customer behavior w.r.t personality traits of agents and customers. The work in [13] examined the relationship between the personality of agents and customer perceptions of service quality; it showed, for example, that openness correlated with assurance, and that conscientiousness predicted reliability. In [3] the authors analyzed which agent traits influence customer satisfaction in different settings (phone, email, on-line chats), and showed that knowledgeableness and preparedness were good indicators. In [17] the authors examined effects of personality traits on customer satisfaction patterns among mobile phone and credit card users. They report that the personality traits modesty, altruism, agreeableness and trust had strong predictive power of customer satisfaction with the two services. There are several differences between these works and ours; first, none of the previous studies considered personality traits of both agents and customers in the interaction. Second, the setting of social media enables new ways for assessments; for example, personality traits in previous work are all based on self report (e.g., IPIP questionnaire [7]), while we suggest automated extraction of traits. This means easier and greater availability and scale of our approach. Third, previous research never considered our unique research goal of optimizing service interactions by recommending a best agent to handle the interaction.

We note that companies like Mattersight[6] provide call center services that match an agent to a customer based on personality traits. Key difference with our approach is that Mattersight matches personality based on caller ID, and previous interactions with the call center. Our approach does not require previous service interactions. We utilize available social profiles and emotions in current interaction.

Works on social media have considered user personality traits [11, 4, 6], but did not focus on the unique context of customers and agents. The focus has been on general issues like engagement in social media, blog topics, discussion topics, etc.

Studies of customer support have documented emotions as part of written interactions. The work in [8], analyzed emotions in textual email communications and focused on prioritizing customer support emails based on detected emotions. Studies of online customer service (chats), such as [18] reported the impact of emotional text usage by service agents on their perception by customers.

3. METHODOLOGY

The objective of our work is to predict customer satisfaction at the end of customer service conversations delivered through social media. We treated this objective as a binary classification task, where the target classes are “satisfied” and “not-satisfied”. The only part of the conversation that is used for this objective is the content of the first message posted by the customer. This means that our classifier generates its prediction as soon as a customer initiated a conversation with the customer service platform, and before a specific agent was assigned to support the customer. We use two auxiliary classifiers to extract affective features. The first auxiliary classifier generates *personality traits* scores based on previous social media posts of the customer and of a possible agent. The second classifier is an *emotion* detection classifier that detects emotions expressed in the customer’s message. These settings enable us to identify and assign an agent, among all available agents, with the personality traits that would maximize the satisfaction of the customer and to prioritize conversations where customers are likely to be dissatisfied.

Below we describe the auxiliary classifiers, the features we extracted from their output and from the customer message content, and the training of our customer satisfaction classifier.

3.1 Personality Traits Classifier

To extract the *personality traits* we utilized the IBM Personality Insights service, available online publicly [7]. This service infers three models of personality, namely, *big five*, *needs* and *values*. The service was trained on social media data, including tweets and forum posts. In total, this classifier extracts percentile scores for 52 traits, as summarized in Table 1. This service requires at least $3,500$ words to have meaningful results. In our collected dataset (described below), 81% of the customers and 91% of the agents answered these requirements. To extract customer *personality traits* we used their historical public tweets as input to the personality traits classifier, and for agents we used their public customer support historical tweets. We distinguished between different agents by parsing their name which appears at the end of each agent tweet in the format of “ˆAGENT_NAME”.

3.2 Emotion Detection Classifier

Another type of affective features we used to predict customer satisfaction is the presence of *emotions* in the content of the first customer message. The emotions are detected by an *emotion* detection classifier based on state-of-the-art features [12, 16, 2], which can detect multiple emotions in each tweet, including: *frustration*, *disappointment*, *confusion*, *politeness* and *anger*.

3.3 Features

We used the following features in our models.

3.3.1 Affective Features

Affective Features comprise two feature families: *personality* and *emotional*. The *personality* family of features are features extracted from *personality traits* of the customer and of the agent assigned to the customer. The extracted features, detailed in Table 2, include the raw output of the *personality traits* classifier for the customer and the agent, and features that represent the interaction of different personality traits of the two parties. These interaction features include root squared error features which capture the similarity between each customer and agent personality trait. MSE and cosine similarity features capture the similarity between a customer and an agent across all personality traits.

[6] http://www.mattersight.com/

[7] https://www.ibm.com/smarterplanet/us/en/ibmwatson/developercloud/doc/personality-insights/index.shtml

Model	Description	Size	Trait
Big five	represents the most widely used model for generally describing how a person engages with the world	35	**Agreeableness** (Altruism, Cooperation, Modesty, Uncompromising, Sympathy, Trust), **Conscientiousness** (Achievement striving, Cautiousness, Dutifulness, Orderliness, Self-discipline, Self-efficacy), **Extraversion** (Activity level, Assertiveness, Cheerfulness, Excitement-seeking, Outgoing, Gregariousness), **Neuroticism** (Fiery, Prone to worry, Melancholy, Immoderation, Self-consciousness, Susceptible to stress), **Openness** (Adventurousness, Artistic interests, Emotionality, Imagination, Intellect, Authority-challenging)
Needs	describes which aspects of a product will resonate with a person	12	Excitement, Harmony, Curiosity, Ideal, Closeness, Self-expression, Liberty, Love, Practicality, Stability, Challenge, Structure
Values	describes motivating factors that influence a person's decision making	5	Self-transcendence, Conservation, Hedonism, Self-enhancement, Excitement

Table 1: Personality traits.

Feature Set Name	# Extracted Features	Mathematical Expression
customer personality traits	52	p_i^c
agent personality traits	52	p_i^a
root squared error	52	$\sqrt{(p_i^c - p_i^a)^2}$
mean squared error (MSE)	1	$\frac{1}{52}\sum_i (p_i^c - p_i^a)^2$
cosine similarity	1	$\frac{\sum_i p_i^c \cdot p_i^a}{\sqrt{\sum_i (p_i^c)^2} \cdot \sqrt{\sum_i (p_i^a)^2}}$

Table 2: Personality feature sets extracted from customer and agent personality traits.

We define p_i^c and p_i^a to be the percentile scores for the customer's and agent's i^{th} personality trait, respectively.

The *emotional* family of features includes the output of the *emotion* detection classifier described above, as a series of binary features (each feature describes a different emotion).

3.3.2 Textual Features

Textual features are extracted from the text of the first customer message, without considering any other information. These features include various n-grams, punctuation and social media features. Namely, *unigrams*, *bigrams*, *NRC lexicon features* (number of terms in a post associated with each affect label in NRC lexicon), and presence of *exclamation marks*, *question marks*, *usernames*, *links*, *happy emoticons*, and *sad emoticons*. These are the features we used in our baseline model detailed in the description of our experiments.

3.4 Customer Satisfaction Prediction System

We trained a binary SVM classifier with a linear kernel. The feature vector used to represent a message incorporated *affective* and *textual* features. A feature vector for a sample in the training data is generated as follows. The *emotion* detection classifier is used on the content of the initial message to output binary emotional scores, that represent whether each emotion is expressed in the content. These scores are then added as the *emotional* features to the feature vector. *Personality* features are generated by running the *personality traits* classifier for the customer and agent, and processing its output to generate the *personality* features described above. *Textual* features are also extracted from the content of the customer message and added to the feature vector. After the model is trained, a test initial message is classified by the model, after being transformed to a feature vector in the same way a train sample is transformed. The SVM classification model outputs a score s where $sign(s)$ determines the class label ("satisfied" or "not-satisfied") while $|s|$ determines the confidence of the classification (which is the distance of the sample from the separating hyper-plane). This can eventually be utilized to assign the most appropriate agent in terms of customer satisfaction confidence. Thus, for a given set of support agents, an agent is assigned such that her personality traits maximize the customer satisfaction confidence.

4. EXPERIMENTS

4.1 Dataset

We gathered data of two North America based Twitter customer service accounts that provide support in English to customers from North America. These dedicated Twitter accounts provide real-time support by monitoring tweets that customers address to it. Corporate support agents reply to these tweets through the Twitter platform. For the two companies, we extracted data from December 2014 until June 2015. For each customer that posted a tweet to the customer support accounts, we searched for the previous message, if any, to which it replied. This allowed us to trace back previous messages and reconstruct the entire conversation. We removed conversations longer than 10 turns, since 89% of the conversations include at most 10 turns, and also removed conversations that contained only 2 messages as these are too short to be meaningful (the customer never replied or provided details about the issue, and thus we can not learn about satisfaction). After applying these preprocessing steps, we had a dataset of $2,632$ conversations.

4.2 Experimental Setup

As a first step to a classification model, we collected ground truth data. For this, we sampled 333 conversations from our dataset, accounting for length frequency in the data set. Each sampled conversation was initiated by a different customer, and the total number of agents in the dataset was 50. We validated that customers and agents in this dataset had enough public tweets available to extract their personality traits. We used Amazon Mechanical Turk[8] to tag the sampled conversations. Each conversation was tagged by five different MTurk's master level judges. Each judge answered the following questions given the full conversation:

- "Overall, how satisfied do you believe the customer was with the service in this communication?"
- "How likely is it that this customer will recommend this service provider to a friend or colleague?"

Each judge indicated an answer on a scale of [0...7], where 0 defines very low agreement, and 7 defines very high agreement with the statement. The average measures of intraclass

[8] https://www.mturk.com/

	Satisfied			Not-Satisfied		
Model	P	R	F	P	R	F
Random	0.721	0.5	0.590	0.279	0.5	0.358
M_t	0.781	0.833	0.806	0.481	0.398	0.435
Aff_{t+p}	0.803	0.833	0.818	0.524	0.473	0.497
Aff_{t+e}	0.809	**0.863**	0.835	0.571	0.473	0.518
Aff_{t+p+e}	**0.827**	0.858	**0.843**	**0.595**	**0.538**	**0.565**

Table 3: Detailed performance results for baseline and affective models.

correlation (ICC) among the judges was 0.847 and 0.845 for the first and the second questions respectively, which indicates high agreement between the judges. Cronbach's Alpha measure for the two questions was 0.942 which indicates very high internal consistency between the questions.

We generated true binary labels for the customer satisfaction classifier from the tagging of the two above mentioned questions. For each conversation, c, we calculated a summary customer satisfaction score, s_c, as the average of the responses to the two questions by each judge, and then averaged the score for all the judges. Using an average of the two ratings rather than only one question is statistically better because it accounts for errors that may have occurred in judges' ratings of one questions, and improves the reliability and validity of the score used [1]. For conversation c, we considered it to end with a positive customer satisfaction if $s_c \geq 4$. This process generated 240 conversations that ended with a positive customer satisfaction and 93 conversations that ended with a negative customer satisfaction.

We evaluated our methods by using leave-one-conversation-out cross-validation as in [10, 9]. Our baseline in all experiments, besides a random classifier, is an SVM classifier that uses only the *textual features* described above, and does not utilize *affective* features. This was used as a state-of-the-art single sentence emotion and sentiment detection approach in many cases (e.g., [12, 16, 15]). Since the classes distribution is unbalanced, we evaluated each class classification performance by using precision (P), recall (R) and F1-score (F). We used Liblinear[9] as an implementation of SVM with a linear kernel and ClearNLP[10] for *textual* features extraction.

4.3 Classification Results

Table 3 depicts the detailed classification results for both classes and a number of models we experimented with. Our baseline models are a model which assigns a label randomly (random) and a model based only on *textual* features (M_t). The novel models we experimented with added *affective* features to the baseline model: a model that uses *textual* and *personality* features (Aff_{t+p}), a model that uses *textual* and *emotional* features (Aff_{t+e}), and a model that uses *textual*, *personality* and *emotional* features (Aff_{t+p+e}). Table 3 shows that all affective models outperform baseline models, where Aff_{t+p} and Aff_{t+e} performed similarly with an average improvement of 17% in F1-score of the "not-satisfied" class, in comparison to M_t. Aff_{t+p+e}, which considered both *emotional* and *personality* affective features, yielded the best performance with an improvement of 30% in F1-score of the "not-satisfied" class, and of 5% for the "satisfied" class. Additionally, we used *McNemar's test* on the contingency tables derived from Aff_{t+p+e} and M_t predictions. This test showed that Aff_{t+p+e} performed statistically significantly better from M_t, under a value of 0.05. These results suggest that utilizing features based on affective components such as *personality traits* and *emotion* expression improves prediction of customer satisfaction to a reasonable level, already after the first message.

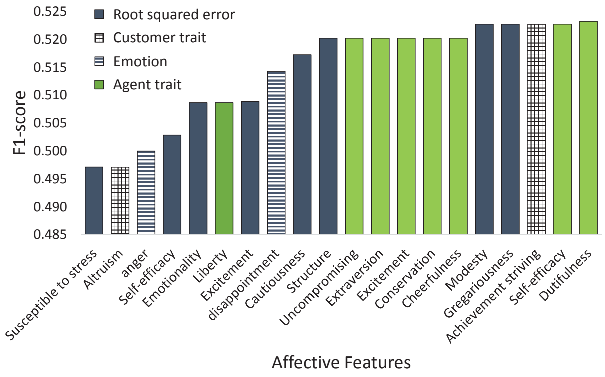

Figure 1: Model performance after excluding specific affective features. The bar's pattern associated with each feature indicates its feature set.

4.4 Contribution of Affective Features

We evaluated the importance of the affective features by performing the following experiment: for each affective feature, f, we generated a model which is trained using all features included in Aff_{t+p+e} except f. For each one of these models we have calculated the F1-score for the "not-satisfied" class. Figure 1 shows 20 affective features for which the lowest F1-scores were obtained, i.e., excluding these features caused performance to deteriorate the most. As Figure 1 shows, removal of each of these 20 features reduced F1 from $F1 = 0.565$ to $F1 < 0.523$. We further see that half of these high effect features describe one of the participants (traits and emotions of customers or agents), and half refer to a fit between them (root squared error). And of the 20 features, only two are emotions (anger and disappointment).

5. CONCLUSIONS AND FUTURE WORK

This paper reports on a first attempt to utilize *affect* features to improve the prediction of customer satisfaction in customer care interactions in social media. We showed how to utilize these features to gain a statistically significant improvement in predicting customer satisfaction. We discussed some practical applications such as optimizing the agent assignment for a specific customer inquiry.

We believe that this is only the tip of the iceberg, and see the following issues as future work. Extend contribution of affective features study, explore the case of virtual agents, extend beyond affect (adding contextual features, such as the topic of the dialogue), experimenting with large-scale datasets (our approach is scalable), and to predict customer satisfaction as the dialogue progresses (not just for the first customer message).

[9] http://liblinear.bwaldvogel.de/

[10] https://github.com/clir/clearnlp

6. ACKNOWLEDGMENTS

The authors thank Michal Jacovi for constructive comments and Ishai Borovoy for assistance in data collection.

7. REFERENCES

[1] E. G. Carmines and R. A. Zeller. *Reliability and validity assessment*, volume 17. Sage publications, 1979.

[2] G. Feigenblat, D. Konopnicki, M. Shmueli-Scheuer, J. Herzig, and H. Shkedi. I understand your frustration. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, pages 25–28. ACM, 2016.

[3] C. M. Froehle. Service personnel, technology, and their interaction in influencing customer satisfaction. *Decision Sciences*, 37(1):5–38, 2006.

[4] H. Gao, J. Mahmud, J. Chen, J. Nichols, and M. X. Zhou. Modeling user attitude toward controversial topics in online social media. In *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014.*, 2014.

[5] K. Gelbrich. Anger, frustration, and helplessness after service failure: coping strategies and effective informational support. *Journal of the Academy of Marketing Science*, 38(5):567–585, 2010.

[6] A. J. Gill, S. Nowson, and J. Oberlander. What are they blogging about? personality, topic and motivation in blogs. In *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM*, 2009.

[7] L. R. Goldberg, J. A. Johnson, H. W. Eber, R. Hogan, M. C. Ashton, C. R. Cloninger, and H. G. Gough. The international personality item pool and the future of public-domain personality measures. *Journal of Research in personality*, 40(1):84–96, 2006.

[8] N. K. Gupta, M. Gilbert, and G. D. Fabbrizio. Emotion detection in email customer care. *Computational Intelligence*, 29(3):489–505, 2013.

[9] E. Ivanovic. Dialogue act tagging for instant messaging chat sessions. In *Proceedings of the ACL Student Research Workshop*, pages 79–84. Association for Computational Linguistics, 2005.

[10] S. N. Kim, L. Cavedon, and T. Baldwin. Classifying dialogue acts in one-on-one live chats. In *Proceedings of EMNLP*, pages 862–871, 2010.

[11] K. Lee, J. Mahmud, J. Chen, M. Zhou, and J. Nichols. Who will retweet this? detecting strangers from twitter to retweet information. *ACM Trans. Intell. Syst. Technol.*, 6(3):31:1–31:25, 2015.

[12] S. Mohammad. Portable features for classifying emotional text. In *Proceedings of NAACL HLT*, pages 587–591, 2012.

[13] L. Neng-Pai, C. Hung-Chang, and H. Yi-Ching. Investigating the relationship between service providers? personality and customers? perceptions of service quality across gender. *Journal of Total Quality Management*, 12(1):57–67, 2001.

[14] R. L. Oliver. *Satisfaction: A behavioral perspective on the consumer*. Routledge, 2014.

[15] A. Qadir and E. Riloff. Learning emotion indicators from tweets: Hashtags, hashtag patterns, and phrases. In *Proceedings of EMNLP*, pages 1203–1209, 2014.

[16] K. Roberts, M. A. Roach, J. Johnson, J. Guthrie, and S. M. Harabagiu. Empatweet: Annotating and detecting emotions on twitter. In *LREC*, pages 3806–3813, 2012.

[17] K. A. Siddiqui. Personality influences on customer satisfaction. *African Journal of Business Management*, 6(11):4134–4141, 2012.

[18] L. Zhang, L. B. Erickson, and H. C. Webb. Effects of emotional text on online customer service chat. In *Graduate Student Research Conference in Hospitality and Tourism*, 2011.