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

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. Result of such an analysis can help personalizing agent assignment such that customer satisfaction is maximized, and prioritizing conversations. In this paper, we show that affective features such as customer's and agent's personality traits and emotional 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 conversation dataset of two customer support services, and shows improvements of 30% in F1-score.

Keywords

ACM proceedings; LaTeX; text tagging

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. In particular, 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 going on 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 service 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 *affects* experienced during a conversation and satisfaction with respect to customer support provided on social media, with the objective of predicting customer satisfaction given affective evaluations of both customers and agents. Specifically we model *affect* by considering both *personality traits* and *emotions*. Those two components capture individuals' intrinsic characteristics with different temporal aspects: personality traits are considered to be permanent while emotions are associated with short durations. 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 [**?**]. As to *personality traits*, many studies examined the effect of specific traits expressed by agents with respect to customer satisfaction, as well as how to interpret traits expressed by customers. For example, traits such as conscientiously, and agreeableness where shown to positively influence the interaction, and traits such as trust and compromising expressed by customers indicate a positive satisfaction [**?**, **?**].

With the advance of behavioral studies in social media, online services are offered for assessing the *personality traits* of social media users based on their online interactions (e.g., tweets, forum posts)[2,3]. These services use linguistic analytics to infer personality and social characteristics. As to *emotions*, analysis services based on textual messages are gaining popularity both in academic studies (cf. [**?**, **?**]), and industry[4,5] as a method to get person valuable insights from 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. Thus, once a message is posted the emotions that were expressed in it are analyzed along with the personality of the customer based on previous interactions in the social media. These data are incorporated together with the personality of the agent that has been assigned for the inquiry to predict the satisfaction of the customer. Companies that

[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/

WOODSTOCK '97 El Paso, Texas USA

DOI: 10.475/123_4

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 that will maximize satisfaction), or to provide the assigned agent information about the behavioral 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

There are various works that studied customers behaviors w.r.t the personality traits of agents and customers. The work in [?] examined the relationship between the personality of the agents and the customer perception on the service quality, and showed for example that openness correlated with assurance, and conscientiousness was a predictor of reliability. In [?] they analyzed which agents traits influence on the customers satisfaction in different settings (phone, email, on-line chats), and showed that knowledgeableness and preparedness are good indicators. In [?] they examined the effects of personality traits on customer satisfaction patterns among mobile phone and credit card users. Their findings were that agreeableness emerged as a single predictor for customer satisfaction for both services. Personality facets modesty, altruism, and trust were consistent in providing major predictive power predicting customer satisfaction for the two services. There are some works in the context of consumers and buying behavior s(rather than customer) including the work in [?], where the authors considered the customer traits in order to improve sales and optimized the consumer demands. Other work that studied the consumers personality is [?]. They found that consumers' characteristic of extraversion is positively related to the level of accrued loyalty, while consumer's characteristic of neuroticism attenuates the association between customer satisfaction and loyalty. Based on those finding they recommended some implications for the stores managers. The main differences between those works to our are as follow; first, none of those studies considered both the personality of the agents as well as the customers in the interaction. Second, the setting of social media open new ways of approaches, for example, the personality traits in those work are all self reported (i.e., filling personality questionnaires like IPIP for big five[6]), while we used automated approach to extract the traits. In addition, the availability and scale of our approach is important. Finally, our unique research goal of optimizing the interaction by recommending on the best agent to handle the interaction was never considered.

It is worth noting, that companies like Mattersight[7] provides service that matches agent to customer based on their personality in call centers. The main differences with our approach is that they match the personality based on the caller id, and her past interactions with the call center, while we consider the current interaction and its emotions and our approach is not obligate to have previous interactions but rather availability of social profile.

In the context of social media, there are works that considered users personality traits such as [?, ?, ?], however they did not focus on the special aspect of customers and agents but rather on general issues like engagement in social media, blogs topics, discussion topics, etc.

In the domain of customer support, several papers studied emotions as part of written interactions. The work in [?], analyzed emotions in textual email communications and the authors focused on prioritizing customer support emails based on detected emotions. In the setting of online customer service (chats), in [?] the authors studied the impact of emotional text on the customer's perception of the service agent.

Emotion detection is also applied to the domain of call centers [?, ?] and this differs from our focus since call center data are voice, and, thus, emotion detection is mainly based on paralinguistic aspects rather than on the text. In addition, if the textual part is considered, then the texts are transcripts of calls that are very different from written text [?], and even more different from the social media setting where the conversation is fully public.

3. METHODOLOGY

The objective of our work is to predict customer satisfaction at the end of customer care conversations in 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 after the customer has initiated the conversation with the customer care service, and before a specific agent was assigned to support the customer. We use two auxiliary classifiers to extract affective features. The first auxiliary classifier is a classifier that generates *personality traits* based on previous social media posts of the customer and the agent (who is possibly assigned to serve the customer). The second classifier is an emotion detection classifier that detects emotions expressed in the customer's message. These settings enable us, among other things, to identify and assign the agent with the personality traits that will maximize the satisfaction of the customer as well as to prioritize conversations.

In this section we describe the auxiliary classifiers, the features we extracted from their output and from the customer message content, and how we trained our "customer satisfaction" classifier. [GF: I think a figure would help here explain the usage of the two classifiers, will work on an option]

3.1 Personality Traits Classifier

To extract the personality traits we utilized the IBM Personality Insights service, publicly available online[8]. Specifically, the 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. Table 1 summarizes the personality traits of the three models. In total classifier extracts scores for 52 traits.The service requires at least 3500 words in order to have meaningful results, and in addition to the percentile scores of the traits, it provides sampling errors.

3.2 Emotion Detection Classifier

[6]https://en.wikipedia.org/wiki/International_Personality_Item_Pool

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

[8]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), Emotional Range (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 characteristics.

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

Figure 1:

Another type of affective features we used to predict customer satisfaction is the presence of emotions, as detected from the content of the first customer message. These emotional features are detected by an emotion detection classifier based on state-of-the-art features [**?**, **?**, **?**]. The classifier can detect multiple emotions in each tweet. The emotion classes that it detects are: *frustration*, *disappointment*, *confusion*, *politeness* and *anger*.

3.3 Features

We used the following features in our models.

3.3.1 Affective Features

Comprises two feature families: *personality* and *emotional*. The *personality* family of features includes features extracted from personality traits of the customer and of the agent who was assigned. The extracted features include the raw output of the personality traits classifier for the customer and the agent, as well as other features that represent the interaction between different personality traits of the two parties (for example, cosine similarity between the customer and the agent traits). We define p_i^c and p_i^a to be the percentile scores for the customer's and agent's ith 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

These features are extracted from the text of the first customer message, without considering any other information. These features include different n-gram, punctuation and social media features. Namely, *unigrams*, *bigrams*, *NRC lexicon features* (number of terms in a post associated with different affect labels in NRC lexicon) *exclamation mark presence*, *question mark presence*, *username presence*, *link presence*, *happy emoticons presence*, and *sad emoticons presence*. We note that 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 we used to represent a message incorporates *affective* and *textual* features. A feature vector for a sample in the train data is generated as follows. The emotion detection classifier is used on the content of the 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 also 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 turn is classified by the model, after 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 be utilized to assign the most appropriate agent in terms of customer satisfaction confidence. I.e., for a given set of support agents assign the agent that its personality traits maximize the customer satisfaction confidence.

4. EXPERIMENTS

4.1 Dataset

We gathered data of two North America based customer support services' Twitter accounts that provide support for customers from North America (so tweets are in English). 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. Specifically, for each customer that posted a tweet to the customer support accounts, we searched for the previous, if any, message to which it replied. Given this method we traced back previous messages and reconstructed the entire conversation. 89% of the conversations include at most 10 turns and thus, we removed conversations longer than 10 turns, and we also removed conversations that contained only 2 messages as these are too short to be meaningful as the customer never replied or provided more details about the issue. After applying these preprocessing steps, we had a dataset of 2632 conversations.

4.2 Experimental Setup

A first step in building a classification model is to obtain ground truth data. For this, we sampled 333 conversations from our dataset, based on their length distribution. Each conversation was initiated by a different customer, and the total number of agents in the dataset was 50. We also validated that each one the customers and agents in this dataset had enough public tweets available to extract their personality traits. The sampled conversations were tagged using Amazon Mechanical Turk[9]. Each conversation was tagged by five different Mechanical Turk'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 her answer on a scale of [0...7], such that 0 defines very low agreement with the statement, and 7 defines a very high agreement with the statement. The intraclass correlation (ICC) among the judges was 0.55 which indicates a moderate agreement.

We generated true binary labels for the customer satisfaction classifier from the tagging of the two above mentioned questions. For each judge we calculated her customer satisfaction tag score as the average of the two answers. For conversation c, we considered it to end with a positive customer satisfaction if $tag(s,c) \geq 4$ where $tag(s,c)$ is the average judges' customer satisfaction tag score, s, for c. 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. 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., [?, ?, ?]). 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[10] as an implementation of SVM with a linear kernel and ClearNLP[11] for textual features extraction.

[9]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+e+p}	**0.827**	0.858	**0.843**	**0.595**	**0.538**	**0.565**

Table 3: Detailed performance results for baseline and affective models

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 *emotional* features (Aff_{t+e}), a model that uses *textual* and *personality* features (Aff_{t+p}) and a model that uses *textual*, *emotional* and *personality* features (Aff_{t+e+p}). 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+e+p}, 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. Additionally, we used *McNemar's test* on the contingency tables derived from Aff_{t+e+p} and M_t predictions. This test showed that Aff_{t+e+p} 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 emotional expression improves prediction of customer satisfaction to a reasonable level, even before the majority of the conversation has taken place.

5. CONCLUSIONS AND FUTURE WORK

This paper reports on a first attempt to utilize affect features (*Personality Traits* and *emotion*) to improve the prediction of customer satisfaction in customer care scenario in social media. We showed how to utilize these features to gain an X% improvement in the customer satisfaction prediction. 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 beyond affect: contextual features, such as the topic of the dialog can add valuable information. For example, it is known that from the customer perspective financial inquires tend to be more sensitive and therefore end with a lower satisfaction [GF: citation].
- Large-scale data: Although in other behavioral studies data sets are even smaller than ours there is an obvious

[10]http://liblinear.bwaldvogel.de/
[11]https://github.com/clir/clearnlp

desire to verify the results on larger data. One of the benefits of our approach is scalability which makes it possible to easily extended to various large data sets on social media platforms.

- Ongoing Dialog: in this work we only consider the first message as the input to predict the satisfaction. However, an interesting setting is to consider the rest of the dialog. As the dialog progresses we can analyze the expressed emotions of the customer and the agent in order to optimize the interaction between the parties with the objective to increase the customer satisfaction. One potential use case can be to escalate the inquiry to a different agent or a supervisor.