Classifying Emotions in Customer Support Dialogues in Social Media

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

Anat Rafaeli, Daniel Altman, David Spivak
Technion-Israel Institute of Technology
Haifa 32000, Israel
Anatr@ie.technion.ac.il,
altmand@campus.technion.ac.il,
dspivak@campus.technion.ac.il

Abstract

Providing customer support through social media channels is gaining increasing popularity. In such a context, automatic detection and analysis of the emotions expressed by customers is important, as is identification of the emotional techniques (e.g., apology, empathy, etc.) in the responses of customer service agents. Result of such an analysis can help assess the quality of such a service, help and inform agents about desirable responses, and help develop automated service agents for social media interactions. In this paper, we show that, in addition to text based turn features, dialogue features can significantly improve detection of emotions in social media customer service dialogues and help predict emotional techniques used by customer service agents.

1 Introduction

An interesting use case for social media is customer support that can now take place over public social media channels. Using this medium has its advantages as described, for example, in (DeMers, 2014): Customers appreciate the simplicity and immediacy of social media conversations, the ability to reach real human beings, the transparency, and the feeling that someone listens to them. Businesses also benefit from the publicity of giving good services almost in real-time, online, building an online community of customers and encouraging more brand mentions in social media. A recent study shows that one in five (23%) customers in the U.S. 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 paper, we analyze customer support dialogues using the Twitter platform and show the utility of such analyses.

The particular aspect of such dialogues that we concentrate on is *emotions*. Emotions are a cardinal aspect of inter-personal communication: they are an implicit or explicit part of essentially any communication, and of particular importance in the setting of customer service, as they relate directly to customer satisfaction and experience (Oliver, 2014). Typical emotions expressed by customers in the context of social media service dialogues include anger and frustration, as well as gratitude and more (Gelbrich, 2010). On the other hand, customer service agents also express emotions in service conversations, for example apology or empathy. However, it is important to note that emotions expressed by service agents are typically governed by company policies that specify which emotions should be expressed in which situation (Rafaeli and Sutton, 1987). This is why we talk in this paper about agent emotional *techniques* rather than agent emotions.

Consider, for example, the real (anonymized) Twitter dialogue depicted in Figure 1. In this dialogue, customer disappointment is expressed in the first turn ('Bummer. =/'), followed by customer support empathy ('Uh oh!'). Then in the last two turns both customer and support express gratitude.

The analysis of emotions being expressed in customer support conversations can take two applications: (1) to discern and compute quality of service indicators and (2) to provide real-time clues to customer service agents regarding the cus-

[1] http://about.americanexpress.com/news/docs/2014x/2014-Global-Customer-Service-Barometer-US.pdf

Proceedings of the SIGDIAL 2016 Conference, pages 64–73,
Los Angeles, USA, 13-15 September 2016.

Figure 1: Example of customer service dialogue that was initiated by a customer (left side), and the agent responses (right side).

tomer emotion expressed in a conversation. A possible application here is recommending to customer service agents what should be their emotional response (for example, in each situation, should they apologize, should they thank the customer, etc.)

Another interesting trend in customer service, in addition to the use of social media described above, is the automation of various functions of customer interaction. Several companies are developing text-based chat agents, typically accessible through corporate web sites, and partially automatized: In these platforms, a computer program handles simple conversations with customers, and more complicated dialogues are transferred to a human agent. Such partially automated systems are also in use for social media dialogues. The automation in such systems helps save human resources and, with further development based on Artificial Intelligence, more automation in customer service chats is likely to appear. Given the importance of emotions in service dialogues, such systems will benefit from the ability to detect (customer) emotions and will need to guide employees (and machines) regarding the right emotional technique in various situations (e.g., apologizing at the right point).

Thus, our goal, in this paper, is to show that the functionality of guiding employees regarding appropriate responses can be developed based on the analysis of textual dialogue data. We show first that it is possible to automatically detect emotions being expressed and, second that it is possible to predict the emotional technique that is likely to be used by a human agent in a given situation. This analysis reflects our ultimate goal: To enable a computer system to discern the emotions expressed by human customers, and to develop computerized tools that mimic the emotional technique used by a human customer service agent in a particular situation.

We see the main contributions of this paper as follows: (1) To our knowledge, this is the first research focusing on automatic analysis of emotions expressed in customer service provided through social media. (2) This is the first research using unique dialogue features (e.g., emotions expressed in previous dialogue turns by the agent and customer, time between dialogue turns) to improve emotion detection. (3) This is the first research studying the prediction of the agent emotional techniques to be used in the response to customer turns.

The rest of this paper is organized as follows. We start by reviewing the related work and a description of the data that we collected. Then we formally define the methodology for detection and prediction of emotion expression in dialogues. Finally, we describe our experiments, evaluate the various models, conclude and suggest future directions.

2 Related Work

2.1 Emotion Detection

Approaches to categorical emotion classification often employ machine learning classifiers, and SVM has typically outperformed other classifiers. In (Mohammad, 2012; Roberts et al., 2012; Qadir and Riloff, 2014) a series of binary SVM classifiers (one for each emotion) were trained over datasets from different domains (news headlines, social media). These works utilize unigrams and bigrams among other lexical based features (e.g., utilizing the NRC emotion lexicon (Mohammad and Turney, 2013)) and punctuation based features. In our work, we also used an SVM classifier, however, while these works aim at classifying single posts (i.e., sentence, tweet, etc.) without context, our work utilizes the context while con-

sidering dialogues. The work in (Hasegawa et al., 2013) showed how to predict and elicit emotions in online dialogues. Their approach for emotion classification is different from ours, for example they only considered the last turn as informative (we consider the full context of the dialogue), and focused on eliciting emotions, while we focus on predicting the agent emotional technique.

2.2 Emotion Expression Prediction

The works in (Skowron, 2010) and (D'Mello et al., 2009) presented dialogue systems that sense the user emotions, such that the system further optimizes its affect response. Both systems use rule-based approaches to generate responses, however, the authors do not discuss how they developed the rules.

It is worth mentioning the works in (Ritter et al., 2011; Sordoni et al., 2015) that are focused on data-driven response generation in the context of dialogues in social media. These works generated general responses, while we focused on predicting the appropriate emotional response.

2.3 Emotions in Written Customer Service Interactions

In the domain of customer support, several papers studied emotions as part of written interactions. The work in (Gupta et al., 2013), 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 (Zhang et al., 2011) the authors studied the impact of emotional text on the customer's perception of the service agent. To extract the emotions, the authors used relatively basic features such as emoticons, exclamation marks, all caps, and some internet acronyms (such as 'lol' or 'imho').

Emotion detection is also applied to the domain of call centers (Vidrascu and Devillers, 2005; Morrison et al., 2007) 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 (Wallace Chafe, 1987), and even more different from the social media setting where the dialogue is fully public.

3 Data

In this section we describe the data collection process and provide some statistics about the Twitter dialogue dataset we have collected.

3.1 Data Collection

Companies that utilize the Twitter platform as a channel for customer service use a dedicated Twitter account which provides real-time support by monitoring tweets that customers address to it. At the same time corporate support agents reply to these tweets also through the Twitter platform. A customer and an agent, can use the Twitter reply mechanism to discuss until the issue is solved (e.g., a solution is provided, or the customer is directed to another channel), or until the customer is no longer active.

In the present work, we define a dialogue to be a sequence of turns between a specific customer and an agent, where the customer initiates the first turn. Consecutive posts of the same party (customer or agent) uninterrupted by the other party, are considered as a single turn (even if there are several tweets). Given the nature of customer support services, we assume the last turn in the dialogue is an agent turn (e.g., "You're very welcome. :) Hit us back any time you need support"). Thus, we expect an even number of turns in the dialogue. We filtered out dialogues in which more than one customer or one agent are involved. Formally, we define a dialogue to be an ordered list of turns $[t_1, t_2, \cdots, t_n]$ where odd turns are customer turns, and even turns are agent turns, and n is even.

Each turn t_i is a tuple consisting of $\{$*turn number*, *timestamp*, *content*$\}$ where *turn number* represents the sequential position of the turn in the dialogue, *timestamp* captures the time the message was published on Twitter, and *content* is the textual message.

3.2 Data Statistics

We gathered data for two North America based customer support services Twitter accounts that provide support for customers from North America (so tweets are in English). One service is for general customer care (denoted as *Gen*), and the other is for technical customer support (denoted as *Tech*). We extracted this 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, turn

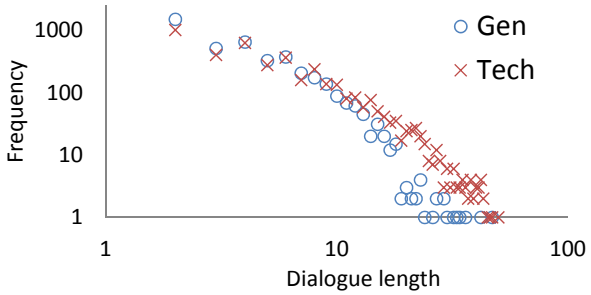

Figure 2: Frequency versus dialogue length for *Gen* and *Tech* on a log-log scale.

	# Dialogues	Mean # turns	AVG word count
Gen	4243	4.83	16.69
Tech	4016	6.81	14.28

Table 1: Descriptive statistics of customer service dialogues extracted from Twitter.

to which it replied. Given this method we traced back previous turns and reconstructed entire dialogues.

Table 1 summarizes some statistics about the collected data, and Figure 2 depicts the frequencies of dialogue lengths which follow a power-law relationship. Table 1 shows differences between the two services; the dialogues in *Tech* tend to be longer (i.e., typically include more turns), with an average of 6.81 turns vs. average of 4.83 turns for *Gen*.

As most of the dialogues include at most 8 turns (88% and 76% for *Gen* and *Tech*, respectively), we removed dialogues longer than 8 turns. In addition, we removed dialogues that contained only 2 turns 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 1189 dialogues of *Gen* support, and 1224 dialogues of *Tech* support.

4 Methodology

The first objective of our work is to detect emotions expressed in customer turns and the second is to predict the emotional technique in agent turns. We treated these two objectives as two classification tasks. We generated a classifier for each task, where the classification output of one classifier can be part of the input to the other classifier. While both classifiers work at the level of turns, i.e., classify the current turn to emotions expressed in it, they are inherently different. When detecting emotions in a customer turn, the turn's content is available at classification time (as well as the history of the dialogue) - meaning, the customer has already provided her input and the system must now understand what is the emotion being expressed. Whereas, when predicting the emotional technique for an agent turn, the turn's content is not available during classification time, but only the agent action and the history of the dialogue since the agent did not respond yet. This difference stems from the fact that in order to train an automated service agent to respond based on customer input, the agent's emotional technique needs to be computed before the agent generates its response sentence.

We defined a different set of relevant emotion classes for each party in the dialogue (customer or agent), based on our above survey of research on customer service (e.g., (Gelbrich, 2010)). Relevant customer emotions to be detected are: *Confusion*, *Frustration*, *Anger*, *Sadness*, *Happiness*, *Hopefulness*, *Disappointment*, *Gratitude*, and *Politeness*. Relevant agent emotional techniques to be predicted are: *Empathy*, *Gratitude*, *Apology*, and *Cheerfulness*.

We utilized the context of the dialogue to extract informative features that we refer to as *dialogue features*. Using these features for emotion classification in written dialogues is novel, and as our experimental results show, it improves performance compared to a model based only on features extracted from the turn's text.

4.1 Features

We used the following features in our models.

4.1.1 Dialogue Features

Comprises three contextual feature families: *integral*, *emotional*, and *temporal*. A feature can be *global*, namely its value is constant across an entire dialogue or it can be a *local*, meaning that its value may change at each turn. In addition, a feature can be *historical* (as will be discussed below).

The *integral* family of features includes three sets of features:

1. *Dialogue topic*: a set of *global* binary features representing the intent of the customer who initiated the support inquiry. Multiple intents can be assigned to a dialogue from a taxonomy of popular topics, which are adapted to the specific service. Examples of topics include *ac-*

count issues, *payments*, *technical problem* and more [2]. This feature set captures the notion that customer emotions are influenced by the event that led the customer to contact the customer service (Steunebrink et al., 2009).

2. *Agent essence*: a set of *local* binary features that represent the action used by the agent to address the last customer turn, independently of any emotional technique expressed. We refer to these actions as the *essence* of the agent turn. Multiple essences can be assigned to an agent turn from a predefined taxonomy. For instance, *"asking for more information"* and *"offering a solution"* are possible essences [3]. This feature set captures the notion that customer emotions are influenced by actions of agents (Little et al., 2013).
3. *Turn number*: a *local* categorical feature representing the number of the turn.

The *emotional* family of features includes *Agent emotion* and *Customer emotion*: these two sets of *local* binary features represent emotions predicted for previous turns. Our model generates predictions of emotions for each customer and agent turn, and uses these predictions as features to classify a later customer or agent turn with emotion expression.

The *temporal* family of features includes the following features extracted from the timeline of the dialogue:

1. *Customer/agent response time*: two *local* features that indicate the time elapsed between the timestamp of the last customer/agent turn and the timestamp of the subsequent turn. This is a categorical feature with values *low*, *medium* or *high* (using categorical values yielded better results than using a continuous value).
2. *Median customer/agent response time*: two *local* categorical features defined as the median of the *customer/agent response times* preceding the current turn. The categories are the same as the previous temporal features.

[2]Currently this feature is not supported in social media. In other channels, for example, customer support on the phone, the customer is requested to provide a topic before she is connected to a support agent (usually using an IVR system). As this feature is inherent in other customer support channels, we assume that in the future it will also be supported in social media.

[3]We assume that if the agent is human, then this input is known to her e.g., based on company policies. For the automated service agent case, we assume that the dialogue system will manage and provide this input.

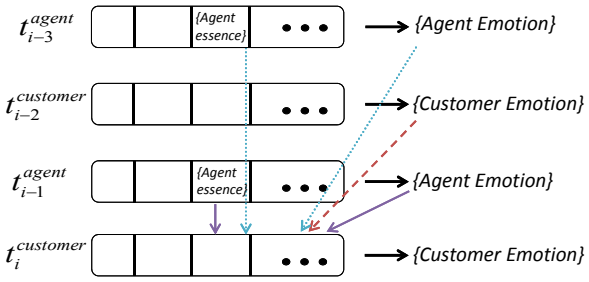

Figure 3: Example for *Historical* features propagation for customer turn, t_i, with $history = 3$. When $history = 1$, the *historical* features are the *agent essence* of turn t_{i-1} and the *agent emotion* predicted for turn t_{i-1} (purple solid line). When $history = 2$, we also add the *customer emotion* detected in turn t_{i-2} (red dashed line). Finally, if we set $history = 3$, then we also add the *agent essence* of turn t_{i-3} and the *agent emotion* predicted for turn t_{i-3} (blue dotted line), so in total we have 5 *historical* features. Notice that the *customer emotion* and *agent essence* features have different values based on their turn number.

3. *Day of week*: a *local* categorical feature indicating the day of the week when the turn was published [Monday - Sunday]. This feature captures the effects of weekend versus weekday influences on emotions (Ryan et al., 2010).

When representing a turn, t_i, as a feature vector, we added some features originating in previous turns $j < i$ to t_i. These features, that are *historical*, include the *emotional* features family and *local integral* features (namely *agent emotions*, *customer emotions* and *agent essence*). We do not include the *turn number* of previous turns, as this is dependent on the turn number of t_i. We denote these features as *historical* features. The value of *history*, that is a parameter of our models, defines the number of sequential turns that precede t_i which propagate *historical* features to t_i.

Figure 3 shows an example of the *historical* features in relation to the classification of customer turn t_i, for *history* size between 1 and 3.

4.1.2 Textual Features

These features are extracted from the text of a customer turn, without considering the context of the dialogue. We use various state-of-the-art text based features that have been shown to be effective for the social media domain (Mohammad, 2012;

Roberts et al., 2012). 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*. We note that these are the features we used in our baseline model detailed below, in the description of our experiments.

4.2 Turn Classification System

For both of the agent and customer turn classification tasks, we implemented two different models which incorporate all of the feature sets we have detailed above. We considered these tasks as multi-label classification tasks. This captures the notion that a party can express multiple emotions (e.g., confusion and anger) in a turn. We chose to use a problem transformation approach which maps the multi-label classification task into several binary classification tasks, one for each emotion class which participates in the multi-label problem (Tsoumakas and Katakis, 2006). For each emotion e, a binary classifier is created using the one-vs.-all approach which classifies a turn as expressing e or not. A test sample is fully classified by aggregating the classification results from all independent binary classifiers. We next define our two modeling approaches.

4.2.1 SVM Dialogue Model

In our first approach we trained an SVM classifier for each emotion class as explained above. The feature vector we used to represent a turn incorporates *dialogue* and *textual features*. The *history* size is also a parameter of this model. Feature extraction for a training/testing feature vector representing a turn t_i, works as follows. *Textual features* are extracted for t_i if it is a customer turn, or for t_{i-1} if it is an agent turn (recall that the system does not have the content of agent turn t_i at classification time). The *temporal* features are also extracted using time lapse values between previous turns as explained above. As discussed above, *agent essence* is assumed to be an input to our module, while *agent emotion* and *customer emotion* features are propagated from classification results of previous turns during testing (or from ground truth labels during training), where the number of previous turns is determined according to the value of *history*. These *historical* features are also appended to the feature vector of t_i, similarly to (Kim et al., 2010) where this method was used for classifying dialogue acts.

4.2.2 SVM-HMM Dialogue Model

Our second approach to classifying dialogue turns is to use a sequence classification method (SVM-HMM), which classifies a sample sequence into its most probable tag sequence. For instance (Kim et al., 2010; Tavafi et al., 2013) used SVM-HMM and Conditional Random Fields for dialogue act classification. Since emotions expressed in customer and agent turns are different, we treated them as different classification tasks (like in our previous approach) and trained a separate classifier for each emotion. We made the following changes when using SVM-HMM:

(1) We treated the emotion classification problem of turn t_i as a sequence classification problem of the sequence $t_1, t_3, ..., t_i$ (i.e., only customer turns) if t_i is a customer turn and $t_2, t_4, ..., t_i$ (i.e., only agent turns) if it is an agent turn. (2) The SVM-HMM classifier generates models that are isomorphic to a k^{th}-order hidden Markov model. Under this model, dependency in past classification results is captured internally by modeling transition probabilities between emotion states. Thus, we removed historical *customer emotion* (resp. *agent emotion*) feature sets when representing a feature vector for a customer (resp. agent) turn. (3) We note that in our setting we provide classifications in real-time during the progress of the dialogue, so at classification time we have access only to previous turns and global information, and we cannot change classification decisions for past turns. Thus, we tagged a test turn, t_i, by classifying the sequence which ends in t_i. Then, t_i was tagged with its sequence classification result.

5 Experiments

5.1 Experimental Setup

A first step in building a classification model is to obtain ground truth data. For this, we sampled dialogues from our dataset, as detailed in Table 2, based on each data source's dialogue length distribution. This sample included 1056 customer turns and 1056 agent turns in total. The sampled dialogues were tagged using Amazon Mechanical Turk[4]. Each dialogue was tagged by five different Mechanical Turk's master level judges. Each

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

Source	# 4 turn dialogues	# 6 turn dialogues	# 8 turn dialogues
Gen	100	66	33
Tech	100	58	38

Table 2: Number of dialogues tagged by judges per source.

judge performed the following tagging tasks given the full dialogue:

1. Emotion tagging: indicate the intensity of emotion expressed in each turn (customer or agent) for each emotion, on a scale of ($[0...5]$), such that 0 defines no emotion, 1 a low emotion intensity and 5 a high emotion intensity. The intraclass correlation (ICC) among the judges was 0.53 which indicates a moderate agreement which is common in this setting (LeBreton and Senter, 2007).
2. Dialogue topic tagging: select one or several topic(s), to represent the customer's intent. The topics are based on a taxonomy of popular customer support topics (Zeithaml et al., 2006): *Account issues, Pricing, Payments, Customer service, Customer experience, Technical problem, Technical question, Order and delivery issues, Behavior of a staff member, Company policy issues* and *General statement*.
3. Agent essence tagging: select one or several of the following for each agent's turn, to describe the agent's action in the specific turn: *Recognizing the issue raised, Asking for more information, Providing an explanation, Offering a solution, General statement* and *Assurance of efforts*. The taxonomy is based on (Zomerdijk and Voss, 2010).

We generated true binary labels from the emotion tagging. For turn t_i, we considered it to express emotion e if $tag(e, t_i) \geq 2$ where $tag(e, t)$ is the average judges' tag value of e in t. This process generated the class sizes detailed in Table 3. Dialogue topic tagging was converted to binary features representing the top-2 selected topics. *Agent essence* feature set representation for each turn was defined analogously. The temporal response time values were translated to *low*/*medium*/*high* categorical values according to their relation to the 33-th and 66-th percentiles.

We evaluated our methods by using leave-one-dialogue-out cross-validation (as in (Kim et al., 2010)), over the whole dataset (for the two customer service data sources together). Each test dialogue was classified by its order of turns, where each turn type (customer or agent) is classified by its corresponding classifier.

Customer		Agent	
Emotion	# of instances	Emotion	# of instances
Happiness	66	Apology	146
Sadness	31	Gratitude	81
Anger	160	Empathy	163
Confusion	68	Cheerfulness	177
Frustration	342		
Disappointment	257		
Gratitude	119		
Hopefulness	30		
Politeness	180		

Table 3: Class size per classification task

Our baseline in all experiments is an SVM classifier that uses only the *textual features* described above, which do not utilize the dialogue context. This was used as a state-of-the-art single sentence emotion detection approach in many cases, e.g., (Mohammad, 2012; Roberts et al., 2012; Qadir and Riloff, 2014) and more. As described above, agent turn emotion prediction is performed before its content is known. Thus, the baseline representation of an agent turn consisted of *textual features* extracted from its preceding customer turn. We evaluated each emotion's classification performance by using precision (P), recall (R) and F1-score (F). We evaluated the total performance for all emotion classes using *micro* and *macro* averages. We used Liblinear[5] as an SVM implementation and SVM-HMM[6] for sequence classification. Additionally, we used ClearNLP[7] for textual features extraction.

5.2 History Size Impact

Since *history* size is a parameter of our models, we first tested the classification results for all possible *history* sizes (given that that maximum dialogue size in our dataset is 8). For each task and for each possible *history* size, we generated *SVM Dialogue* and *SVM-HMM Dialogue* models and evaluated them as detailed above. We compared the *macro* and *micro* average *F1-score* of our classifiers against the baseline classifier performance. As depicted in Figure 4 both the *SVM Dialogue* and *SVM-HMM Dialogue* models were superior

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

[6] https://www.cs.cornell.edu/people/tj/svm_light/svm_hmm.html

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

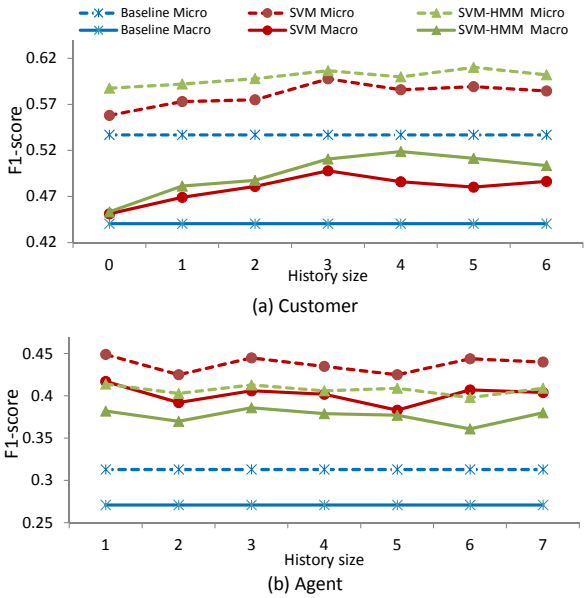

Figure 4: Macro and micro average F1-score for various history sizes for customer (a) and agent (b) turn classifiers.

for all history ranges and for both tasks. Examining the customer turns emotion detection performance, we can see in Figure 4(a) that it increases until $history = 3$, and then remains relatively stable for larger *history* sizes. This means that information about the behavior of the customer and agent in past turns is beneficial for detecting customer emotions in a current turn. For assessing the performance of our predictions of agent turns emotion techniques, we first note that we tested with $history > 0$ range, since we assume that the minimal information needed for agent turn classification is the information extracted from the last customer turn. Figure 4(b) shows that overall, performance is highest when $history = 1$, and does not decline much for higher *history* values. This indicates that for agent emotion technique prediction the last customer turn is the most informative one.

In all of our experiments, we used the *Wilcoxon signed-rank test* to validate the statistical significance of our models' *micro* and *macro* average *F1-score* comparing to baseline performance. Additionally, we used *McNemar's test* on the contingency tables aggregated over all emotions. These tests showed that both of our models were significantly different from the baseline model, under a value of 0.001, for both classification tasks and all *history* sizes.

5.3 Detailed Classification Results

Table 4 depicts the detailed classification results for optimal $history$ values that obtained maximal *macro F1-score*, namely for customer emotion detection $history = 4$ and for agent emotion technique prediction $history = 1$. The table presents performance for each emotion, for *macro* and *micro* average results over all dialogues, and for each data source (*Gen* or *Tech*) separately. For both classification tasks, both of our models outperformed baseline results for almost all emotions, where average *macro* and *micro* results are statistically significant compared to the baseline, as described above.

For customer turn emotion detection, the *SVM-HMM Dialogue* model performed better than the *SVM Dialogue* model, and reached a *macro* and *micro* average *F1-score* improvements over all dialogues of 17.8% and 11.7%, respectively. Furthermore, the *macro* and *micro* average *F1-score* results of the *SVM-HMM Dialogue* model (0.519 and 0.6, respectively) are satisfying given the moderate ICC score between the judges (0.53). For predicting the agent emotional technique, the *SVM Dialogue* model obtained slightly better results than *SVM-HMM Dialogue* model, and reached a *macro* and *micro* average *F1-score* improvements over all dialogues of 53.9% and 43.5%, respectively. These results emphasize the differences between the *SVM Dialogue* and *SVM-HMM Dialogue* models. Specifically, when $history$ size is large, as in customer emotion prediction, *SVM-HMM Dialogue* model, which internally captures dependencies in past classifications, outperforms the simplistic *SVM Dialogue* model. We note that an improvement is also obtained when calculating *macro* and *micro* average performance for each data source separately. This highlights our models' superiority as well as their general applicability and robustness for different data sources.

5.4 Feature Set Contribution Analysis

We examined the contribution of different feature sets in an incremental fashion, using the optimal *history* value detailed above. Based on the families of feature sets that we defined in the Methodology section, we tested the performance of different feature set combinations in our models, added in the following order: *baseline* (textual features), *emotional*, *temporal* and *integral*. Figure 5 depicts

Classification task	Emotion	Baseline			SVM Dialogue Model				SVM-HMM Dialogue Model			
		P	R	F	P	R	F	%	P	R	F	%
Customer emotion detection	Happiness	.556	.379	.450	.622	.424	.505	12.0	**.627**	**.561**	**.592**	31.4
	Sadness	.412	.226	.292	.429	.194	.267	-8.6	**.444**	**.258**	**.327**	12.0
	Anger	.615	.469	.532	**.669**	.569	.615	15.6	.638	**.606**	**.622**	16.9
	Confusion	.200	.147	.169	**.255**	.191	.218	28.9	.254	**.221**	**.236**	39.4
	Frustration	**.667**	.608	.636	.659	.623	.641	.7	.659	**.673**	**.666**	4.7
	Disappointment	.529	.432	.475	.618	**.572**	**.594**	24.9	**.628**	.553	.588	23.7
	Gratitude	.786	.739	.762	**.827**	**.765**	**.795**	4.3	.826	.756	.789	3.6
	Hopefulness	.133	.067	.089	**.286**	.067	.108	21.6	.280	**.233**	**.255**	186.4
	Politeness	.607	.472	.531	**.618**	.494	.549	3.4	.561	**.583**	**.572**	7.7
	Gen - macro	.540	.405	.463	.582	.456	.511	10.3	**.592**	**.514**	**.551**	18.9
	Gen - micro	.685	.527	.596	**.716**	.606	.657	10.2	.691	**.641**	**.665**	11.6
	Tech - macro	.394	.332	.361	**.478**	.356	.408	13.2	.457	**.419**	**.437**	21.3
	Tech - micro	.450	.410	.429	**.482**	.417	.447	4.2	.479	**.469**	**.474**	10.5
	Total - macro	.500	.393	.440	**.554**	.433	.486	10.4	.546	**.494**	**.519**	17.8
	Total - micro	.597	.488	.537	**.637**	.543	.586	9.1	.617	**.583**	**.600**	11.7
Agent emotional technique prediction	Apology	.276	.264	.270	.418	**.423**	**.420**	55.6	**.424**	.380	.400	48.1
	Gratitude	.108	.049	.068	**.326**	**.197**	**.245**	260.3	.200	**.197**	.198	191.2
	Empathy	.287	.240	.261	**.401**	**.390**	**.395**	51.3	**.401**	.349	.373	42.9
	Cheerfulness	.491	.463	.477	**.592**	**.598**	**.594**	24.5	.546	.564	.554	16.1
	Gen - macro	.310	.275	.291	**.488**	**.462**	**.474**	62.9	.450	.433	.441	51.5
	Gen - micro	.342	.281	.308	**.489**	**.468**	**.478**	55.2	.461	.429	.444	44.2
	Tech - macro	.216	.201	.208	**.277**	**.263**	**.269**	29.3	.265	.256	.260	25.0
	Tech - micro	.338	.302	.319	**.425**	**.392**	**.407**	27.6	.379	.366	.372	16.6
	Total - macro	.290	.254	.271	**.434**	**.402**	**.417**	53.9	.393	.372	.382	41.0
	Total - micro	.340	.289	.313	**.463**	**.437**	**.449**	43.5	.427	.403	.414	32.3

Table 4: Detailed performance results for customer and agent classification tasks given optimal *history* size. For brevity, the table presents improvement relative to baseline in percentages only for *F1-score*.

the results for both classification tasks. The x-axis represents specific combination of features sets, and the y-axis represents the *macro* or *micro* average *F1-score* value obtained. Figure 5 shows that adding each feature set improved performance for all models, for both tasks, which indicates the informative value of each feature set. Additionally, the figure suggests that the most informative dialogue feature sets are the *integral* and *emotional*.

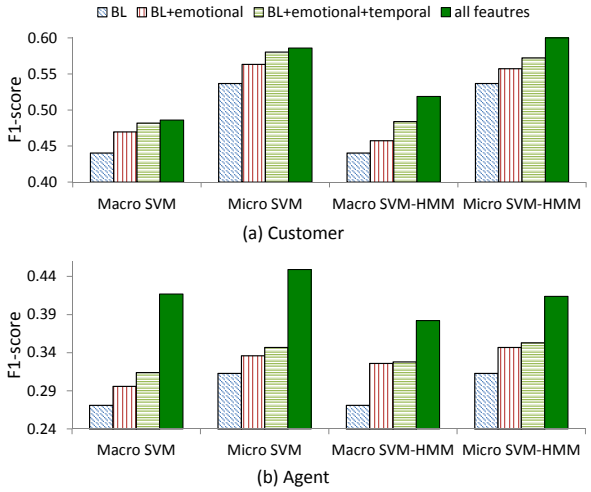

Figure 5: Macro and micro average F1-score for various feature set combinations for customer (a) and agent (b) turn classifiers. BL stands for baseline.

6 Conclusions

In this work we studied emotions being expressed in customer service dialogues in the social media. Specifically, we described two classification tasks, one for detecting customer emotions and the other for predicting the emotional technique used by support service agent. We have proposed two different models (*SVM Dialogue* and *SVM-HMM Dialogue* models) for these tasks. We studied the impact of *dialogue features* and dialogue *history* on the quality of the classification and showed improvement in performance for both models and both classification tasks. We also showed the robustness of our models across different data sources. As for future work we plan to work on several aspects: (1) In this work, we showed that it is possible to predict the emotional technique. In the future, we plan to run experiments in which the predicted emotional technique is actually applied in the context of new dialogues to measure the effect of such predictions on real support dialogues. (2) Distinguish between dialogues that have positive outcomes (e.g., high customer satisfaction) and others.

References

Jayson DeMers. 2014. 7 reasons you need to be using social media as your customer service portal. *Forbes*.

Sidney D'Mello, Scotty Craig, Karl Fike, and Arthur Graesser. 2009. Responding to learners' cognitive-affective states with supportive and shakeup dialogues. In *Proceedings of HCI*, pages 595–604.

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

Narendra K. Gupta, Mazin Gilbert, and Giuseppe Di Fabbrizio. 2013. Emotion detection in email customer care. *Computational Intelligence*, 29(3):489–505.

Takayuki Hasegawa, Naoki Yoshinaga Kaji, Nobuhiro and, and Masashi Toyoda. 2013. Predicting and eliciting addressee's emotion in online dialogue. In *ACL (1)*, pages 964–972.

Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. 2010. Classifying dialogue acts in one-on-one live chats. In *Proceedings of EMNLP*, pages 862–871.

James M LeBreton and Jenell L Senter. 2007. Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*.

Laura M Little, Don Kluemper, Debra L Nelson, and Andrew Ward. 2013. More than happy to help? customer-focused emotion management strategies. *Personnel Psychology*, 66(1):261–286.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

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

"Donn Morrison, Ruili Wang, and Liyanage C. De Silva. 2007. Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication*, 49(2):98–112.

Richard L Oliver. 2014. *Satisfaction: A behavioral perspective on the consumer*. Routledge.

Ashequl Qadir and Ellen Riloff. 2014. Learning emotion indicators from tweets: Hashtags, hashtag patterns, and phrases. In *Proceedings of EMNLP*, pages 1203–1209.

Anat Rafaeli and Robert I Sutton. 1987. Expression of emotion as part of the work role. *Academy of management review*, 12(1):23–37.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of EMNLP*.

Kirk Roberts, Michael A Roach, Joseph Johnson, Josh Guthrie, and Sanda M Harabagiu. 2012. Empatweet: Annotating and detecting emotions on twitter. In *LREC*, pages 3806–3813.

Richard M Ryan, Jessey H Bernstein, and Kirk Warren Brown. 2010. Weekends, work, and well-being: Psychological need satisfactions and day of the week effects on mood, vitality, and physical symptoms. *Journal of social and clinical psychology*, 29(1):95–122.

Marcin Skowron. 2010. Affect listeners: Acquisition of affective states by means of conversational systems. In *Development of Multimodal Interfaces: Active Listening and Synchrony*, pages 169–181.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Meg Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *In NAACL-HLT*.

B.R. Steunebrink, M.M. Dastani, and J.-J.Ch. Meyer. 2009. The occ model revisited. In *Proceedings of the 4th Workshop on Emotion and Computing*, pages 478–484.

Maryam Tavafi, Yashar Mehdad, Shafiq Joty, Giuseppe Carenini, and Raymond Ng. 2013. Dialogue act recognition in synchronous and asynchronous conversations. In *Proceedings of the SIGDIAL*, pages 117–121.

Grigorios Tsoumakas and Ioannis Katakis. 2006. Multi-label classification: An overview. *Dept. of Informatics, Aristotle University of Thessaloniki, Greece*.

Laurence Vidrascu and Laurence Devillers. 2005. Detection of real-life emotions in call centers. In *INTERSPEECH*, pages 1841–1844.

Deborah Tannen Wallace Chafe. 1987. The relation between written and spoken language. *Annual Review of Anthropology*, pages 383–407.

Valarie A Zeithaml, Mary Jo Bitner, and Dwayne D Gremler. 2006. Services marketing: Integrating customer focus across the firm.

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

Leonieke G Zomerdijk and Christopher A Voss. 2010. Service design for experience-centric services. *Journal of Service Research*, 13(1):67–82.