Linguistic Elements of Engaging Customer Service Discourse on Social Media

Sonam Singh[1] **and Anthony Rios**[2]
[1]Department of Marketing
[2]Department of Information Systems and Cyber Security
University of Texas at San Antonio
{sonam.singh, anthony.rios}@utsa

Abstract

Customers are rapidly turning to social media for customer support. While brand agents on these platforms are motivated and well-intentioned to help and engage with customers, their efforts are often ignored if their initial response to the customer does not match a specific tone, style, or topic the customer is aiming to receive. The length of a conversation can reflect the effort and quality of the initial response made by a brand toward collaborating and helping consumers, even when the overall sentiment of the conversation might not be very positive. Thus, through this study, we aim to bridge this critical gap in the existing literature by analyzing language's content and stylistic aspects such as expressed empathy, psycho-linguistic features, dialogue tags, and metrics for quantifying personalization of the utterances that can influence the engagement of an interaction. This paper demonstrates that we can predict engagement using initial customer and brand posts.

1 Introduction

Providing quality customer service on social media has become a priority for most companies (brands) today. A simple customer-brand interaction started by a moment of annoyance can be relieved when the brand resolves the issue in a public display of exceptional customer service. According to Forbes, companies that use Twitter as a social care channel have seen a 19 percent increase in customer satisfaction (Forbes, 2018). Furthermore, customers are rapidly turning to social media for customer support; Research from JD Power finds that approximately 67 percent of consumers now tap networks like Twitter and Facebook for customer service (Power, 2013). While providing timely and stellar service has its advantages, engaging in a collaborative dialogue with its customers also leads to mutual trust and transparency according to social customer relations management (SCRM) theories (Yahav et al., 2020). SCRM has been documented as a core business strategy (Woodcock et al., 2011). SCRM refers to "*a philosophy and a business strategy, supported by a technology platform, business rules, processes, and social characteristics, designed to engage the customer in a collaborative conversation to provide mutually beneficial value in a trusted and transparent business environment*" (Greenberg, 2010). While brand agents on social platforms are typically motivated and well-intentioned to help engage with customers, their efforts are often ignored if their initial response to the customer does not match a specific tone, style, or topic the customer is aiming to receive. Hence, we explore what textual elements of a brand's response are predictive of customer engagement.

While there has been substantial research on customer service on social platforms, a majority has predominantly addressed issues such as timely response (Xu et al., 2017), timely transfer from a bot to human (Liu et al., 2020), improving bot performance (Adam et al., 2020; Følstad and Taylor, 2019; Xu et al., 2017; Hu et al., 2018), improving dialogue act prediction (Oraby et al., 2017; Bhuiyan et al., 2018), and managing customer sentiment of users on the platform (Mousavi et al., 2020). A few studies have also examined the tone and emotional content of the customer service chats (Hu et al., 2018; Herzig et al., 2016; Oraby et al., 2017). However, more subtle and integral stylistic aspects of language, such as expressed empathy, psycho-linguistic features (e.g., time orientation), and the level of personalization of the responses, have received little attention, particularly when analyzed for engagement metrics. For the few studies that have looked at subtle stylistic aspects of language in customer service settings (Clark et al., 2013; Wieseke et al., 2012), the studies were more lab-based in synchronous, face-to-face, or call settings and may not translate to asynchronous, text-based contexts. Additionally, only emotional

Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 105 - 117

aspects of empathy were considered. Yet, similar to Face-to-Face communication, text-only communication also contains many subtle, and not so subtle, social cues. (Jacobson, 1999; Hancock and Dunham, 2001; Bargh and McKenna, 2004; Rouse and Haas, 2003). Broadly, there are two dimensions of language that can provide information about a conversation: language *content* and *style*. The *content* of a conversation indicates the general topics being discussed, along with the relevant actors, objects, and actions mentioned in the text. Conversational *style* reflects how it is said (Pennebaker, 2011). Thus, a text-based response can be examined for its content and its style.

Language can be viewed as a fingerprint or signature (Pennebaker, 2011). Besides reflecting information about the people, organizations, or the society that created it, the text also impacts the attitudes, behavior, and choices of the audience that consume it (Berger et al., 2020). For example, the language of the response from a brand agent can assure and calm a consumer, infuriate them, or make a customer anxious. While language certainly reflects something about that writer (e.g., their personality, how they felt that day, and how they feel towards someone or something), the language also impacts the people who receive it (Packard and Berger, 2021; Packard et al., 2018). It can influence customer attitudes toward the brand, influence future purchases, or affect what customers share about the interaction (Berger et al., 2020).

Overall, in this paper, we aim to examine how the initial query from a customer and the initial response from a brand's agent impact the engagement of interaction on social media. However, rather than focusing just on the textual content of a conversation alone, we also examine the language style through the use of cognitive and emotional expressed empathy (e.g., emotional reactions, interpretations, explorations) (Sharma et al., 2020), psycho-linguistic (LIWC) language features (e.g., time orientation, tone) (Pennebaker et al., 2015),dialogue-tags, and novel use of perplexity as a metric of personalization of the responses (Brown et al., 1992; Heafield, 2011). Overall, to the best of our knowledge, we make a first attempt at examining a comprehensive set of content and style features from customer service conversations to understand their impact on *engaging conversations* between brand agents and customers. In addition, this is the first study to analyze customer engagement as a measure of the effort of brand agents to provide customer support beyond positive sentiment. Moreover, we also build a prediction model to demonstrate the predictive capability of these stylistic features. We show it is possible to predict the likelihood of its engagement from the first response from a brand agent.

2 Related Work

Given the importance of customer service there exists a substantial body of research addressing different challenges in this area. A body of researchers focuses on improving chatbots. Adam et al. (2020) build upon social response theory and anthropomorphic design cues. They find artificial agents can be the source of persuasive messages. However, the degree to which humans comply with artificial social agents depends on the techniques applied during human-chatbot communication. In contrast, Xu et al. (2017) designed a new customer service chatbot system that outperformed traditional information retrieval system approaches based on both human judgments and an automatic evaluation metric. Hu et al. (2018) examine the role of tone and find that tone-aware chatbot generates as appropriate responses to customer requests as human agents. More importantly, the tone-aware chatbot is perceived to be even more empathetic than human agents. Følstad and Taylor (2019) emphasize the repair mechanisms (methods to fix bad chatbot responses) and find that chat-bots expressing uncertainty are bad in the customer service setting. Thus, Følstad and Taylor (2019) develop a method to suggest likely alternatives in cases where confidence falls below a certain threshold.

Another stream of research examines the role of emotions and sentiment in service responses. For example, Zhang et al. (2011) find that emotional text positively impacts customers' perceptions of service agents. Service agents who use emotional text during an online service encounter were perceived to be more social. Guercini et al. (2014) identify elements of customer service-related discussions that provide positive experiences to customers in the context of the airline industry. They find that positive sentiments were linked mostly to online and mobile check-in services, favorable prices, and flight experiences. Negative sentiments revealed problems with the usability of companies' websites, flight delays, and lost luggage. Evidence of delightful experiences was recorded among ser-

vices provided in airport lounges. Mousavi et al. (2020) explores the factors and external events that can influence the effectiveness of customer care. They focus on the telecom industry and find that the quality of digital customer care that customers expect varies across brands. For instance, customers of higher priced firms (e.g., Verizon and AT&T) expect better customer care. Different Firms provide different levels of customer service and seemingly unrelated events (e.g., signing an exclusive contract with a celebrity) can impact digital customer care.

There is also research that focuses on "dialogue acts"—identifying utterances in a dialogue that perform a customer service-specific function. For instance, Herzig et al. (2016) study how agent responses should be tailored to the detected emotional response in customers, in order to improve the quality of service agents can provide. They demonstrate that dialogue features (e.g., dialogue acts/topics) can significantly improve the detection of emotions in social media customer service dialogues and help predict emotional techniques used by customer service agents. Similarly, Bhuiyan et al. (2018) develop a novel method for dialogue act identification in customer service conversations, finding that handling negation leads to better performance. Oraby et al. (2017) use 800 twitter conversations and develop a taxonomy of fine-grained "dialogue acts" frequently observed in customer service. Example dialogue acts include, but are not limited to, complaints, requests for information, and apologies. Moreover, Oraby et al. (2017) show that dialogue act patterns are predictive of customer service interaction outcomes. While their outcome analysis is similar to our study (i.e., predicting successful conversations), it differs in the ultimate analysis and goals. Specifically, they focus on function rather than language style. For example, their work can guide service representatives to ask a yes-no question, provide a statement, or ask for information. In contrast, in this paper, we focus on looking at specific language features. Our work can guide *how* the customer representative should respond (e.g., provide a more specific response), which is different than describing *what* they should do (e.g., ask for more information).

Finally, closely related to our study, there are a range of studies examining the different aspects of language and its impact on customer service. For example, Packard and Berger (2021) study linguistic correctness—the tangibility, specificity, or imaginability of words employees use when speaking to customers. They find customers are more satisfied, willing to purchase and purchase more when employees speak to them concretely. Clark et al. (2013) study the nature and value of empathetic communication in call center dyads. They find that affective expressions, such as "*I'm sorry*," were less effectual, but attentive and cognitive responses could cause highly positive responses, although the customers' need for them varied substantially. Wieseke et al. (2012) find that customer empathy strengthens the positive effect of employee empathy on customer satisfaction, leading to more "symbiotic interactions." The major difference between the prior studies and ours is that they study empathy in lab-based settings (i.e., real-world interactions) and not text conversations on social media. Furthermore, we correlate language to consumer engagement, which is missing in prior work.

	Train	**Test**
Total Conversations	472,412	317,348
Engaging	134,650	44,796
Average Convo. Length	4.15	4.14
Max Convo. Length	604[1]	432

Table 1: Dataset Statistics

3 Data

Our data set consists of customer service-related queries and brand responses from their Twitter service handles. We start with two million tweets (2,013,577) spanning over 789K conversations between 108 brands and 667,738 customers. Some brands have as few tweets as 107 conversations (e.g., OfficeSupport), while other brands such as Amazon and Apple have 100K and 36K conversations, respectively. Figure 1 plots the brands that have more than 10K tweets. The average number of tweets per conversation is 4.15 and the average number of words per tweet is 18.52. The length of these conversations varies substantially, while some are as short as one round (user tweet→brand tweet→end) others are as long as ten rounds. We measure *engagement* by counting the number of brand→customer interactions are made. For instance, if the customer writes a question, a brand responds, and then the customer never responds again, then that would have an engagement count of zero. If the customer responds once, then the engagement count is one.

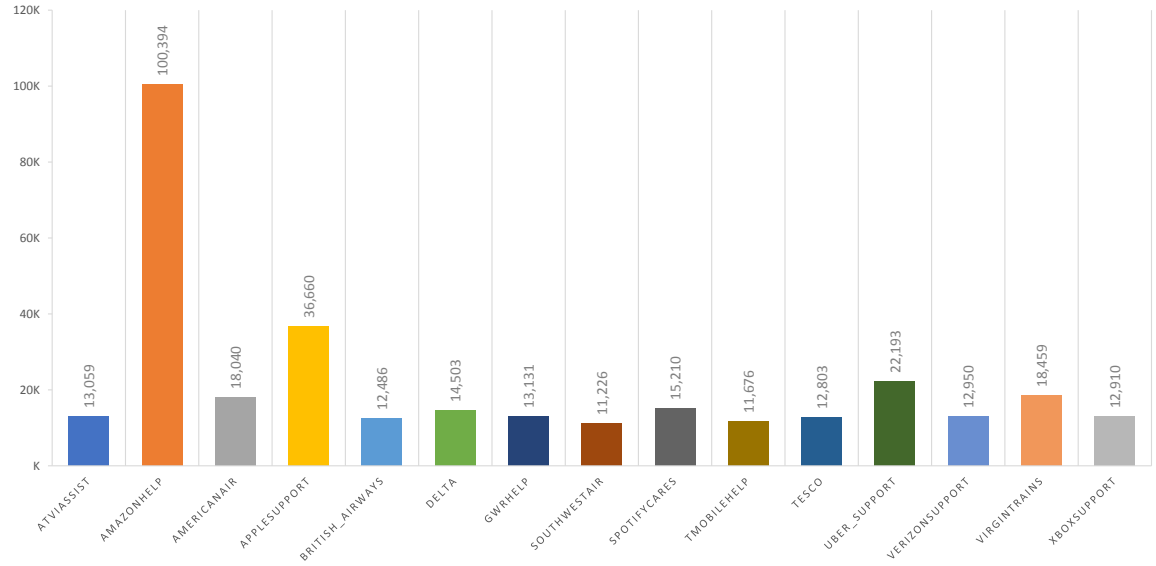


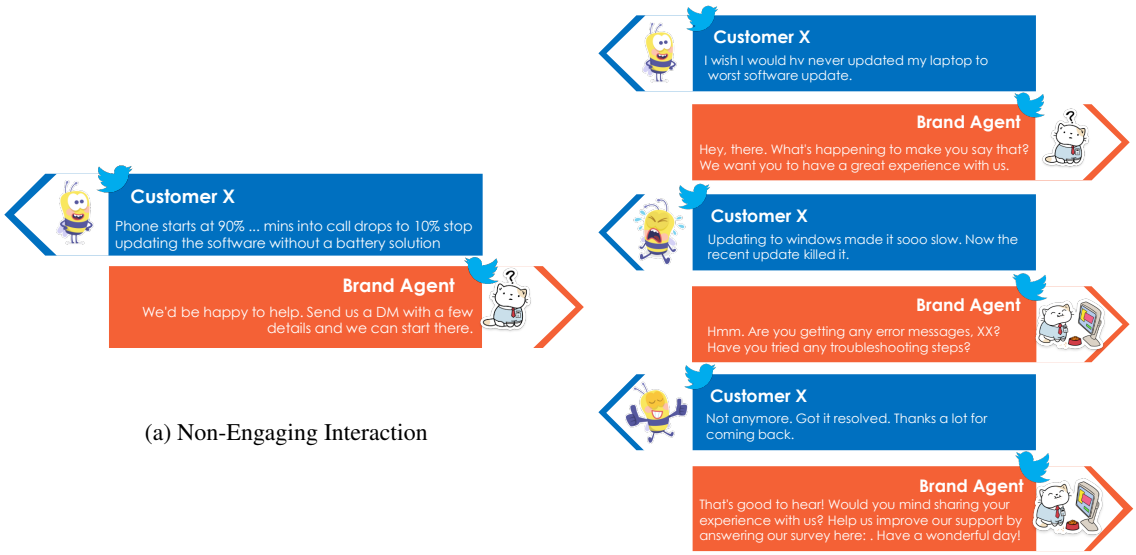
Figure 1: Plot of the number of tweets for Brands with at least 10K tweets in our dataset.

To understand engagement better, we manually analyzed 500 conversations with more than 1 round (user→brand→user...) to understand the nature of these conversations. We found that 85% of these conversations were putting substantial personalized effort towards trying to resolve customer issues, 1% were appreciations from customers, and the remaining 14% represented customers' frustrations/anger with a poor experience. Even among the 14% of the conversations that expressed anger initially, we found that the brand agent showed effort towards helping the customer when the customer actively engaged with them. Thus, we find that the length of a conversation can reflect the effort and quality of the initial response made by a brand toward collaborating and helping consumers. While many of these conversations might get resolved offline and reflect neutral sentiment, or limited engagement, on social platforms, the conversation size can provide a helpful signal in finding conversation-related characteristics that signal quality brand responses. Hence, we are able to use this metric of engagement to better understand which language factors lead to it. We operationalize the engagement in two forms. First, we consider the length of the conversation, for example, Figure 2b has a length of three. The length of the conversation represents the effort brand agents put in to resolve customer query. Second, we also construct a binary "Engagement Indicator" outcome variable which is "engaging" when the discourse has more than one round (user→brand→user→..), and "not-engaging" when it ends in one round (user→brand→end). Figure 2b and Figure 2a provide a sample example of a engaging vs. not engaging conversation. Table 1 provides a summary of the dataset statistics.

[1]Note that this includes multi-party conversations, the original user only replied four times in the longest conversation

4 Method

Our main research question is to understand how the content and stylistic features of the text influence the engagement of interaction in social media discourse. Thus, our methodology involves five major steps. In Step 1, we collect and understand data to identify engagement through the length. For Step 2, we generate the stylistic metrics to cover empathy, personalization/novelty of a response, dialogue tags, and general psycho-linguistic features. We also include content-based features. For Step 3, we operationalize engagement as an Engagement Indicator ("engaged", if the length is greater than one or "not engaged" otherwise). Step 4 involves machine learning experiments. We start with all two million tweets (2,013,577) spanning over 789K conversations between 108 brands and 667,738 customers. We then randomly split (60:40) the entire data set into training and test sets (train size- 472,512; test size- 317,348). For Step 5 we report the most predictive items for all features. The details of the feature generation process are described in the following subsections.



(a) Non-Engaging Interaction

(b) Engaging Interaction

Figure 2: Example of engaging and non-engaging interactions. The non-engaging interaction has a single customer response (length = 1) and the engaging conversation has multiple customer responses (length = 3).

4.1 Content Features

We explore two types of content features: bag-of-word features and neural content features generated using RoBERTa (Liu et al., 2019). We describe both sets of features below:

Bag-of-Words. We use TF-IDF-weighted unigrams and bigrams (Schütze et al., 2008) to build our content features for both customer and the brand's response posts. Specifically, we make use of content features from both the **Customer Post** and **Brand Agent Post**. We experiment with using either the initial customer or brand posts independently. Likewise, we evaluate the performance of using both posts. Note, when combing the posts, we treat the features from each group independently, e.g., there are two features for the word "great," one for the customer post and one for the brand post.

RoBERTA. We also experiment with the RoBERTa model (Liu et al., 2019). Ideally, we hope that our engineered features can match the performance of a complex neural network-based method, while also resulting in an interpretable model. Hence, we compare with RoBERTa which is a strong baseline for many text classification tasks. Specifically, we experiment with using both the initial customer post $P = [w_1, \ldots, w_N]$ and the initial brand post $B = [w_1, \ldots, w_T]$ where w_i represents word i. We evaluate the performance of each independently, along with combining them where we concatenate the brand post to the end of the customer post before processing with RoBERTa. The last layer's CLS token is passed to a final output layer for prediction.

4.2 Stylistic Features

As previously mentioned, in this study, we analyze both content and stylistic features. Content features measure what the text is about. Style features measure how it is written. This can include information about its author, its purpose, feelings it is meant to evoke, and more (Argamon et al., 2007). Hence, to construct the stylistic features of the discourse, we examine the expressed empathy of the response posts, the psycho-linguistic features of both customer and brand response posts, the dialogue tags, and the perplexity (uniqueness) of the brand's response posts. Each set of features is described below.

Empathy Identification Model For empathy identification, we use the framework introduced by Sharma et al. (2020). It consists of three communication mechanisms providing a comprehensive outlook of empathy—*Emotional Reactions*, *Interpretations*, and *Explorations*. *Emotional reactions* involve detecting texts related to emotions such as warmth, compassion, and concern, expressed by the brand agent after hearing about the customers' issue (e.g., *I'm sorry you are having this problem*). *Interpretations* involve the brand agent communicating a real understanding of the feel-

Tags	Examples
Statement	Updating to windows made it sooo slow. Now the recent update killed it.
Question	Is this something you're seeing now? Let us know in a DM and we'll take it from there.
Appreciation	Thanks I've since had an email from XXX and it's been sorted.
Response	Not a problem XX, I hope your future journeys are better. ZZZ.
Suggestion	That's good to hear! Would you mind sharing your experience with us? Help us improve our support by answering our survey here: Have a wonderful day!

Table 2: Sample dialogue tags.

	BC3		QC3	
	BERT	JM	BERT	JM
Accuracy	.89	.85	.87	.78
Macro F1	.67	.63	.70	.65

Table 3: Comparison of BERT to Joty and Mohiuddin (2018) on the BC3 and QC3 dialogue acts datasets.

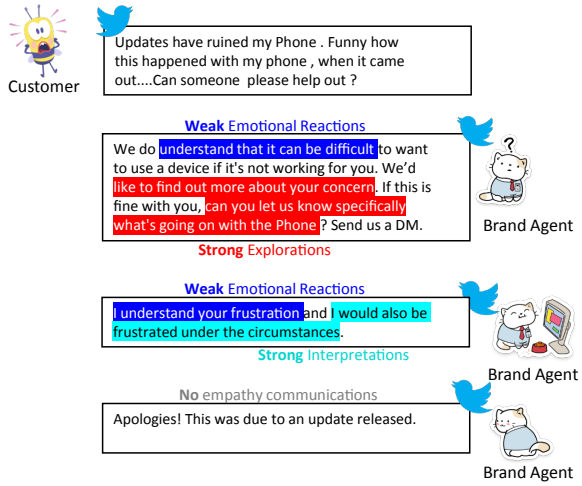


Figure 3: Examples of the three empathy communication mechanisms: Emotional Reactions, Interpretations, and Explorations based upon (Sharma et al., 2020). We differentiate between no communication, weak communication, and strong communication of these factors.

ings and issues of the customer (e.g., *I have also had t his issue before, I'm sorry, it really is annoying*). *Explorations* involve the brand agent actively seeking/exploring the experience of the customer (e.g., *what happened when you restarted your computer?*). For each of these mechanisms, the study differentiates between, (0) no expression of empathy (no communication), (1) weak expression of empathy (weak communication), (2) strong expression of empathy (strong communication). To get these scores for each empathy mechanism, we use the pre-trained RoBERTa-based model by Sharma et al. (2020). The model leverages attention between neural representations of the seeker and response posts for generating a seeker-context aware representation of the response post, used to perform the two tasks of empathy identification and rationale extraction. We leverage this model and train it on the Reddit corpus of the Sharma et al. (2020) dataset for 4 epochs using a learning rate of 2e-5, batch size of 32, λEI $= 1$, and λRE $= .5$. Figure 3 provides sample responses to represent the three mechanisms of empathy. Someone can argue about the bias the training data set might introduce, given it is not related to customer support. However, we believe that the context is quite similar and there are several similarities that should minimize any bias. First, both data sets have a similar structure. A seeker who is seeking some answer or issue resolution that has been bothering him from an agent responsible for providing the response. Second, both data sets are online text-based communications, thus minimizing the bias again. We believe "Empathy" is a very universal thing and should not differ much with text-based communications. To evaluate this, we sampled a small set of tweets from our dataset and qualitatively found the model reliable. The examples provided in Figure 3 are slight variants of tweets within our dataset that were correctly identified.

Perplexity. How can we measure whether a brand agent's response is generic or not? To do this, we use custom-built language models and measure their perplexity on each tweet. Specifically, we train a KenLM (Heafield, 2011) n-gram-based language model on a held-out set of responses across all brands. We then use the language model to calculate the probability of a response generated by the agent. We use the perplexity metric to score the response (probabilities), which is a commonly used metric for measuring the performance of a language model. Perplexity is defined as the inverse

of the probability of the test set normalized by the number of words

$$PPL(X) = \sqrt{\prod_{i=1}^{N} P(w_i|w_{i-1})^{-\frac{1}{N}}}$$

where $P(w_i|w_{i-1})$ represents the probability of a word given the previous word and N is the total number of words in a given agent's response. The equation above is an example using only ngrams. We train the KenLM model with a combination of 3-, 4-, and 5-grams.

Intuitively, another way of interpreting perplexity is the measure of the likelihood of a given test sentence in reference to the training corpus. Based on this intuition, we hypothesize the following: "When a language model is primed with a collection of response tweets, the perplexity can serve as an indicator for personalization of a given brand's response." The rationale behind this is that the most common tweets would share more similarities (e.g., common terms and language patterns) with each other. This leads to common responses such "*How may I help you?*" to have lower perplexity while unique responses such as "*Sorry to hear that! What is the exact version of the OS you're running and we'll figure out our next steps there. Thanks.*" will have a higher perplexity score. This hypothesis is supported by the use of perplexity to measure "surprisal" of misinformation and fake news when primed on factual knowledge (Lee et al., 2021).

Dialogue Tags. When people interact on social media, they interact with each other at different times, performing certain communicative acts, called speech acts (e.g., question, request, statement). We hypothesize that the types of communicative acts made by the user and brand agent can have an impact on overall engagement. Table 2 provides examples for dialogue tags. To perform deep conversational analysis, we fine-tune a transformer model BERT (Devlin et al., 2018) on the QC3 (Joty and Mohiuddin, 2018) and BC3 (Ulrich et al., 2008) data sets for speech act recognition and achieve superior performance than original study (Joty and Mohiuddin, 2018). We then use our model for qc3—based on the overall performance of their respective datasets and simple qualitative analysis of our data—to score dialogue tags for the initial posts for both customers and brand agents. Since, the speech acts identify the sentence structure as a question, suggestion, statement, appreciation, or response, which is more specific to linguistic structure than domain, we believe the bias introduced through the training data set would be minimal. Table 3 compares the results for our model. The predicted tags for each item is tweet is used as a feature in our final model.

Model	Macro P.	Macro R.	Macro F1
Stratefied Baseline	.50	.50	.50
Uniform Baseline	.50	.50	.41
Minor Class baseline	.06	.50	.10
RoBERTa Models			
RoBERTa + Customer Post	.59	.57	.58
RoBERTa + BAP	.73	.72	.72
RoBERTa + CP + BAP	**.73**	.73	**.73**
Linear BoW Models			
CP	.58	.61	.58
BAP	.65	.77	.67
CP + BAP	.65	.75	.68
LIWC + E + P + DT	.57	.68	.54
CP+BAP+LIWC+E+P+DT	.69	**.76**	**.72**

Table 4: Main Results for different feature sets: Customer Post (CP), Brand Agent Post (BAP), Perplexity (P), Dialogue Tags (DT), and Empathy (E).

	Macro P.	Macro R.	Macro F1
CP + BAP + LIWC + E + P+ DA	.69	.77	.72
– perplexity (P)	.60	.73	.57
– empathy (E)	.60	.73	.58
– LIWC	.65	.77	.66
– Dialogue Acts (DA)	.66	.78	.69
– Brand Agent Post (BAP)	.62	.65	.63
– Customer Post (CP)	.67	.80	.69

Table 5: Ablation Results for different feature sets using the linear model: Customer Post (CP), Brand Agent Post (BAP), Perplexity (P), Dialogue Tags (DT), and Empathy (E).

Psycho-Linguistic Features. To examine the language more deeply, we leverage the psycholinguistic resources from the Linguistic Inquiry and Word Count (LIWC) tool (Pennebaker et al., 2001, 2007) which has been psycho-metrically validated and performs well on social media data sets to extract lexico-syntactic features (De Choudhury et al., 2013). LIWC provides a set of lexicons (word lists) for studying the various emotional, cognitive, and structural components present in individuals' written text. We extract several linguistic measures from LIWC, including the word count, psychological, cognitive, perceptual processes, and time orientation separately for customer and brand posts. We run each post independently through LIWC to generate independent scores.

5 Experiments

This section evaluates how style and content features impact the general predictive performance of the machine learning models.

Model Training Details. For classification, we used the dichotomous dependent variable: "Engagement Indicator". Specifically, using the features described in the previous section, we train a classification model. The classification model is trained to predict the "engagement" class, i.e., whether the length of the customer→brand responses is greater than one. We train the Logistic regression model using the scikit-learn package (Pedregosa et al., 2011). Additionally, we also trained a transformer-based model-RoBERTa with binary cross-entropy classification loss. We trained this model using a learning rate of 2e-5 and a batch size of 32 on 4 epochs. Hyperparameters for all models were chosen using a randomly sampled validation partition of 10% of the training data.

Baselines. We report the results of various baselines. Specifically, we compare various naive baselines, including three random classification baselines: Stratified, Minor Class, and Uniform for classification. Stratified randomly generates predictions based on the class (Engaging and Not-Engaging) proportions. Uniform randomly generates predictions equally for both classes, independent of the class frequency. The Minor Class baseline always predicts the least frequent class. Beyond the naive baselines, we compare three models using content features: Logistic Regression models trained on the Customer Post, Brand Agent Post, and Customer + Brand Agent Posts. All models using content features make use of TF-IDF-weighted unigram and bigram features. Furthermore, we compare to a model that uses the stylistic features for customers and brand agents (LIWC + Empathy + Perplexity + Dialogue Tags), both independently and combined. Finally, "our" method uses all of the content and style features across brands and customers (Customer Post + Brand Agent Post + LIWC + Empathy + Perplexity + Dialogue Tags).

Results. Table 4 reports the Macro Precision, Macro Recall, and Macro F1. We make two major observations. First, we find that our method outperforms the naive (random) baselines. The Macro-F1 score for the naive baselines is highest for the Stratified Baseline with an F1 of .50. On the contrary, the Minor class Baseline, performs poorly for the Macro F1 score (.10), i.e., because it always predicts the "non-engagement" class. Next, we find that Brand Post content features are more predictive than the customer's original post. Hence, the Brand's response is vital for promoting engagement, more so than the original customer's tweet. Finally, the combination of content features and stylistic features performs best with a Macro of .72 almost matching the best Macro F1 (0.73) for the more complex (uninterpretable) RoBERTa model. Moreover, the combination method outperforms all other methods with regard to Macro Recall, suggesting that the linear models are more robust using all of the engineered features. See the Appendix for implications and further discussion about the results.

Ablation Study. Next, we analyze the components in our classification models through an ablation study for the model using all the features. Intuitively, we wish to test which set of features has the largest impact on model performance. Table 5 summarizes our findings. Interestingly, we see the most significant drops in performance from removing Perplexity and Empathy information (e.g., removing perplexity features drops the Macro F1 from .72 to .57, and removing empathy drops it to .58), indicating complex relationships with the other features in the classification model.

Feature Importance. Next, we perform a comprehensive analysis of our model focusing on the coefficient scores of the logistic regression model to analyze individual feature impact on model performance. Our paper reveals several insights for brand agents as well as customers. Table 6 summarizes our feature importance results for features with the largest magnitudes (positive and negative). At a high level, we find that Empathy Explorations are of substantial importance for positive engagement. Likewise, the LIWC category Clout indicates negative relationships for engagement. This is interesting because a Higher Clout score is marked by using more we-words and social words and fewer I-words and negations (e.g., no, not). This indicates that users engage more when brands take responsibility for issues (e.g., "I will find you a solution" vs. "we can work together to fix it"). This is further supported by the positive relationship for words with high certainty made used by the Brand (i.e., BRAND: certain). Lastly, we find that Novelty

Feature Group	Feature	Importance
Empathy	Exploratorations	.126
	Interpretations	-.004
	Emotional Reactions	-.034
Dialogue Tags	BRAND: questions	.072
	CUSTOMER: statements	.019
	CUSTOMER: response	.012
	BRAND: suggestions	.007
	CUSTOMER: appreciations	-.003
	CUSTOMER: suggestions	-.006
	CUSTOMER: questions	-.019
	BRAND: statements	-.042
	BRAND: appreciations	-.044
	BRAND: response	-.049
LIWC	CUSTOMER: word_count	.136
	BRAND: word_count	.116
	BRAND: interrogation	.092
	CUSTOMER: time	.089
	BRAND: time	.088
	BRAND: focuspast	.061
	BRAND: focuspresent	.041
	BRAND: certain	.040
	CUSTOMER: tentative	-.014
	BRAND: informal	-.015
	CUSTOMER: informal	-.016
	CUSTOMER: focusfuture	-.017
	CUSTOMER: focuspast	-.035
	BRAND: focusfuture	-.055
	BRAND: insight	-.114
	BRAND: Tone	-.139
	BRAND: Clout	-.319
Personalization	BRAND: Novelty	-.026

Table 6: Feature Importance for Stylistic Features used to predict engagement

has a small negative coefficient score. After further analysis, we find that when there is a slight chance that a highly novel initial response by the brand can quickly solve the problem right away, limiting the need for further discussion—which is a good thing. However, this is somewhat rare in our analysis. The overall recommendations based on our findings are summarized in Figure 4. We summarize the key implications/recommendations in the following subsections.

6 Conclusion

Our study demonstrates that even though some customer support requests on social media might reflect anger, there are some key indicators that can engage customers in a positive direction. Most extant research has not paid any significant attention to this aspect of the discourse, mostly simply focusing on sentiment or timely response. Hence, we examined text based, asynchronous social media discourses for both *what is written* and *how it is written*, to examine how these features influence customer engagement. Our study effectively identifies multiple such stylistic features that can influence the engagement of these social discourses between customers and brands.

7 Limitations and Future Research

A limitation of our study is that we proxy engagement with the length or the number of rounds customers and brands respond to each other on the social platform and assume that customers appreciate the continuous effort from the brand to resolve their issues, even when they are not resolved. Our initial analysis supports this, however, future research can validate the perceived effort through a natural field experiment or a lab setting. Moreover, we do not account for engagement outside the social platform. Another limitation of our paper is that it is exploratory and based on observational data. A controlled lab experiment studying the sentiment of the conversations along with the stylistic/linguistic features to contrast the two aspects and support the claim empirically can establish the claims further. Furthermore, even when this is not the complete case, understanding what gets responses is important, even if it is to point agents towards what to avoid. Also, while we believe engagement is a proxy for quality interactions from the customers' perspective (in general), engagement is not a good thing in all cases from a customer service perspective because it increases the time agents are working with each customer on average. However, these results are also useful for future research in customer service chatbot development, which can be useful to develop bots that show direct care for customers. Moreover, while it can increase the cost of customer service, social media is also acting as a strong marketing tool, which can increase revenue for the company, hence, the limitation may not be as strong as potentially expected; however, this would need to be tested.

Acknowledgements

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 2145357.

References

Martin Adam, Michael Wessel, and Alexander Benlian. 2020. Ai-based chatbots in customer service and

their effects on user compliance. *Electronic Markets*, pages 1–19.

Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6):802–822.

John A Bargh and Katelyn YA McKenna. 2004. The internet and social life. *Annu. Rev. Psychol.*, 55:573–590.

Jonah Berger, Ashlee Humphreys, Stephan Ludwig, Wendy W Moe, Oded Netzer, and David A Schweidel. 2020. Uniting the tribes: Using text for marketing insight. *Journal of Marketing*, 84(1):1–25.

Mansurul Bhuiyan, Amita Misra, Saurabh Tripathy, Jalal Mahmud, and Rama Akkiraju. 2018. Don't get lost in negation: An effective negation handled dialogue acts prediction algorithm for twitter customer service conversations. *arXiv preprint arXiv:1807.06107*.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, Jennifer C Lai, and Robert L Mercer. 1992. An estimate of an upper bound for the entropy of english. *Computational Linguistics*, 18(1):31–40.

Colin Mackinnon Clark, Ulrike Marianne Murfett, Priscilla S Rogers, and Soon Ang. 2013. Is empathy effective for customer service? evidence from call center interactions. *Journal of Business and Technical Communication*, 27(2):123–153.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of ICWSM*, volume 7.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Asbjørn Følstad and Cameron Taylor. 2019. Conversational repair in chatbots for customer service: The effect of expressing uncertainty and suggesting alternatives. In *International Workshop on Chatbot Research and Design*, pages 201–214. Springer.

Forbes. 2018. Forbes. https://www.forbes.com/sites/shephyken/2018/08/05/what-customers-want-and-expect/#5a4c6f7a7701. Accessed: 2021-01-01.

Paul Greenberg. 2010. *CRM at the speed of light: Social CRM strategies, tools, and techniques*. McGraw-Hill New York.

Simone Guercini, Fotis Misopoulos, Miljana Mitic, Alexandros Kapoulas, and Christos Karapiperis. 2014. Uncovering customer service experiences with twitter: the case of airline industry. *Management Decision*.

Jeffrey T Hancock and Philip J Dunham. 2001. Impression formation in computer-mediated communication revisited: An analysis of the breadth and intensity of impressions. *Communication research*, 28(3):325–347.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197.

Jonathan Herzig, Guy Feigenblat, Michal Shmueli-Scheuer, David Konopnicki, Anat Rafaeli, Daniel Altman, and David Spivak. 2016. Classifying emotions in customer support dialogues in social media. In *Proceedings of SIGdial*, pages 64–73, Los Angeles. Association for Computational Linguistics.

Tianran Hu, Anbang Xu, Zhe Liu, Quanzeng You, Yufan Guo, Vibha Sinha, Jiebo Luo, and Rama Akkiraju. 2018. Touch your heart: A tone-aware chatbot for customer care on social media. In *Proceedings of CHI*, pages 1–12.

David Jacobson. 1999. Impression formation in cyberspace: Online expectations and offline experiences in text-based virtual communities. *Journal of Computer-Mediated Communication*, 5(1):JCMC511.

Shafiq Joty and Tasnim Mohiuddin. 2018. Modeling speech acts in asynchronous conversations: A neural-crf approach. *Computational Linguistics*, 44(4):859–894.

Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2021. Towards few-shot fact-checking via perplexity. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1971–1981.

Jiawei Liu, Zhe Gao, Yangyang Kang, Zhuoren Jiang, Guoxiu He, Changlong Sun, Xiaozhong Liu, and Wei Lu. 2020. Time to transfer: Predicting and evaluating machine-human chatting handoff. *arXiv preprint arXiv:2012.07610*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Reza Mousavi, Monica Johar, and Vijay S Mookerjee. 2020. The voice of the customer: Managing customer care in twitter. *Information Systems Research*, 31(2):340–360.

Shereen Oraby, Pritam Gundecha, Jalal Mahmud, Mansurul Bhuiyan, and Rama Akkiraju. 2017. " how may i help you?" modeling twitter customer service conversations using fine-grained dialogue acts. In *Proceedings of the IUI*, pages 343–355.

Grant Packard and Jonah Berger. 2021. How concrete language shapes customer satisfaction. *Journal of Consumer Research*, 47(5):787–806.

Grant Packard, Sarah G Moore, and Brent McFerran. 2018. (i'm) happy to help (you): The impact of personal pronoun use in customer–firm interactions. *Journal of Marketing Research*, 55(4):541–555.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

James W Pennebaker. 2011. Using computer analyses to identify language style and aggressive intent: The secret life of function words. *Dynamics of Asymmetric Conflict*, 4(2):92–102.

James W Pennebaker, Roger J Booth, and Martha E Francis. 2007. Linguistic inquiry and word count: Liwc [computer software]. *Austin, TX: liwc. net*, 135.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Power. 2013. Power 2013. https://www.jdpower.com/business/press-releases/2013-social-media-benchmark-study. Accessed: 2021-01-01.

Steven V Rouse and Heather A Haas. 2003. Exploring the accuracies and inaccuracies of personality perception following internet-mediated communication. *Journal of research in personality*, 37(5):446–467.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*.

Jan Ulrich, Gabriel Murray, and Giuseppe Carenini. 2008. A publicly available annotated corpus for supervised email summarization.

Jan Wieseke, Anja Geigenmüller, and Florian Kraus. 2012. On the role of empathy in customer-employee interactions. *Journal of service research*, 15(3):316–331.

Neil Woodcock, Andrew Green, and Michael Starkey. 2011. Social crm as a business strategy. *Journal of Database Marketing & Customer Strategy Management*, 18(1):50–64.

Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A new chatbot for customer service on social media. In *Proceedings of CHI*, pages 3506–3510.

Inbal Yahav, David G Schwartz, and Yaara Welcman. 2020. The journey to engaged customer community: Evidential social crm maturity model in twitter. *Applied Stochastic Models in Business and Industry*, 36(3):397–416.

Lu Zhang, Lee B Erickson, and Heidi C Webb. 2011. Effects of "emotional text" on online customer service chat. *Graduate Student Research Conference in Hospitality and Tourism*.

A Appendix

A.1 Brand Agent Implications

We examine both the emotional and cognitive aspects of expressed empathy and find that it can positively affect the likelihood of a successful interaction. Similarly, the level of personalization or novelty (measured using perplexity) of the brand agent's initial response and their time orientation also reveal some interesting insights for brand managers. The findings of our study provide new insights to brands for orchestrating an effective and engaging customer service discourse on social platforms, thus building great customer relationships. Additionally, our findings can also provide insights into improving customer service chatbot responses by incorporating the stylistic features that we discuss in the study. We summarize some of the main findings that are relevant for brands below:

Expressing more *exploration* empathy about customer's issue in the initial response increases the likelihood of an engaging interaction. We find that the fact whether or not an agent's response communicates emotional (e.g., *I'm sorry*) reactions might not be as important as explorations (i.e., *can you share the error message on your iPhone?*) for an engaging conversation indicating that exploring issues that a customer is facing influences the engagement of an interaction. Moreover, indicating generic understanding in form of interpretation empathy might help to resolve the issue quickly as it affects the length of the conversation negatively. This is also validated by other sets of stylistic features - Dialogue Tags (BRAND: questions and BRAND: suggestions) and LIWC

(BRAND: interrogation) when brand agents ask more questions and provide more suggestions that lead to more engaging conversations by eliciting responses from customers.

Initial responses focused on the future from brand agent decreases the engagement of an interaction Brand agents often use phrases such as "*We will look into this*" or "*We will get back to you*". Our results show that the higher future orientation (*LIWC-BRAND: focusfuture*) in the initial response post can lower the engagement of an interaction. An alternative solution based on our results could be to use exploration empathy to understand more about the customer's issue. On the other hand, the use of past-tense verbs ("*I fixed it for you*") and present focused ("*We are looking into this now"* increase the engagement of an interaction. Thus, a general recommendation is to avoid making future promises, instead focus on responding when the incident is resolved, or state that you are actively working on it.

A.2 Customers/Users Implications

Our study also identifies some key implications for customers. The main intention of any customer reaching out to a brand on social media most typically is to get some issue resolved quickly. No matter how motivated or well-intentioned brand agents might be to resolve these issues, given the frequency and load of such issues they might leave many of such posts unattended or not provide satisfactory resolutions. We therefore also provide some guidelines to customers on how to increase the effectiveness of such interactions on social platforms.

Interrogative customer posts with informal language lower the engagement of an interaction Our results show that the more interrogative and informal the initial customer post is (*CUSTOMER: questions, CUSTOMER: informal*), the less likely it is going to be engaging. For instance, customer posts asking questions and using swear words or informal language are less likely to be engaging on the platform. This finding is substantial as this can help customers to frame their issues in a more explanatory manner rather than being informal and interrogative of the brands.

Customer posts focused too much in the past or future or tentative are less engaging. We find that when the initial customer posts contain past-focused or future-focused words such as 'had it enough' 'will see you, or "may" or "perhaps it is likely to lead to an engaging interaction. This is an interesting finding because the usage of such words might signal that customer is not expecting their issue to be resolved and thus, brands might choose to attend to other posts, given the rate of customer service requests pouring on social media. For instance, given two customer posts "*May be someone could help with this issue?*" and "*Please help to resolve this issue*", the latter post signals the brand agent to take an action (and increases the likelihood of a success), while the former that signals tentative action.

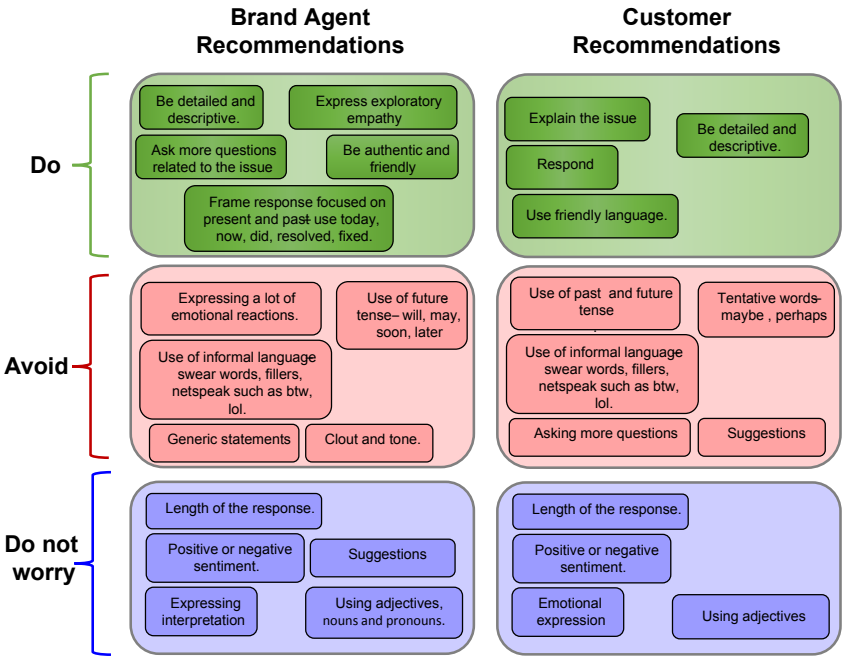


Figure 4: Overall recommendations for brand agents and customers based on this paper's findings.