A Million Tweets Are Worth a Few Points:
Tuning Transformers for Customer Service Tasks

Amir Hadifar[†]   **Sofie Labat**[‡]   **Veronique Hoste**[‡]
Chris Develder[†]   **Thomas Demeester**[†]

[†] IDLab, Dept. of Information Technology, Ghent University - imec, Belgium
[‡] LT3, Dept. of Translation, Interpreting and Communication, Ghent University, Belgium

Abstract

In online domain-specific customer service applications, many companies struggle to deploy advanced NLP models successfully, due to the limited availability of and noise in their datasets. While prior research demonstrated the potential of migrating large open-domain pretrained models for domain-specific tasks, the appropriate (pre)training strategies have not yet been rigorously evaluated in such social media customer service settings, especially under multilingual conditions. We address this gap by (i) collecting a multilingual social media corpus containing customer service conversations (865k tweets), (ii) comparing various pipelines of pretraining and finetuning approaches, (iii) applying them on 5 different end tasks. We show that pretraining a generic multilingual transformer model on our in-domain dataset, before finetuning on specific end tasks, consistently boosts performance, especially in non-English settings.[1]

1 Introduction

Online platforms and social media are increasingly important as communication channels in various companies' customer relationship management (CRM). To ensure effective, qualitative and timely customer service, Natural Language Processing (NLP) can assist by providing insights to optimize customer interactions, but also in real-time tasks: (i) detect emotions (Gupta et al., 2010), (ii) categorize or prioritize customer tickets (Molino et al., 2018), (iii) aid in virtual assistants through natural language understanding and/or generation (Cui et al., 2017), etc.

Despite this NLP progress for CRM, often small and medium-sized companies (SMEs) struggle with applying such recent technology due to the limited size, noise and imbalance in their datasets. General solutions to such challenges are transfer learning strategies (Ruder, 2019): *feature extraction* uses frozen model parameters after pretraining on an external (larger) training corpus, while *finetuning* continues training on the smaller in-domain corpus. In the large body of work adopting such strategies (e.g., Pan and Yang 2009), little effort has been put into addressing specific CRM use cases that need to rely on social media data that is noisy, possibly multilingual, and domain-specific for a given company.

In this paper, we analyze the possibilities and limitations of transfer learning for a number of CRM tasks, following up on the findings of Gururangan et al. (2020) who demonstrate gains from progressive finetuning on in-domain and task-specific monolingual data. Specifically, our contributions are that we (1) collect a multilingual corpus of 275k Twitter conversations, comprising 865k tweets between customers and companies in 4 languages (EN, FR, DE, NL), covering distinct sectors (telecom, public transport, airline) (Section 4.1); (2) rigorously compare combinations of pretraining and finetuning strategies (Section 3) on 5 different CRM tasks (Section 4.2), including prediction of complaints, churn, subjectivity, relevance, and polarity; and (3) provide empirical results (Section 5). We find that additional pretraining on a moderately sized in-domain corpus, before task-specific finetuning, contributes to overcoming the lack of a large multilingual domain-specific language model. Its effect is much stronger than consecutive finetuning on smaller datasets for related end tasks. Furthermore, our experimental results show that when pretrained models are used in feature extraction mode, they struggle to beat well-tuned classical baselines.

2 Related Work

A wide range of NLP research has been devoted to customer services. Hui and Jha (2000) employed data mining techniques to extract features from a

---

[1]Dataset and code available at https://github.com/hadifar/customerservicetasks

customer service database for decision support and machine fault diagnosis. Gupta (2011) extracted a set of sentiment and syntactic features from tweets for customer problem identification tasks. Molino et al. (2018) introduced the Customer Obsession Ticket Assistant for ticket resolution, using feature engineering techniques and encoder-decoder models. Highly popular pretrained language models, such as BERT (Devlin et al., 2019), have also been explored for different customer service tasks: Hardalov et al. (2019) considered re-ranking candidate answers in chatbots, while Deng et al. (2020) proposed BERT-based topic prediction for incoming customer requests.

Although the performance gains obtained by pretraining language models are well-established, they need further exploration in terms of multilinguality. Some studies (Pires et al., 2019; Karthikeyan et al., 2019; Wu et al., 2019) have investigated the transferability of multilingual models on different tasks, but they do not consider the effect of progressive pretraining on a smaller and less diverse multilingual corpus, as we will do.

3 Methodology

3.1 Architecture

We selected some of the most popular publicly available pretrained language models to explore transfer learning properties for CRM classification tasks: RoBERTa (Liu et al., 2019), XLM (Conneau et al., 2020), and BERTweet (Nguyen et al., 2020). These models are pretrained on the English Wikipedia and BookCorpus (Zhu et al., 2015), CommonCrawl in 100 languages, and 850M English tweets, respectively. The XLM and BERTweet pretraining procedure is based on RoBERTa, which itself is a transformer-based Masked Language Model (MLM; Devlin et al., 2019). All of these models require a different classifier 'head' for each target task to estimate the probability of a class label.

3.2 Transfer Strategies

We adopt a straightforward approach to transfer learned representations: we continue pretraining the considered transformer models on a 4-lingual corpus of customer service Twitter conversations (see Section 4.1), i.e., the overall domain of all considered sub-tasks. After that, we apply additional adaptation for cross-lingual transfer (Section 5.1), as well as cross-task transfer (Section 5.2).

The following notations are used throughout the rest of this paper to describe pretraining stages:
- π – further **p**retraining the original MLM on our 4-lingual tweet conversation corpus.
- φ – **fi**netuning the pretrained model extended with the MLP classifier on the target task
- $\varnothing$ – freezing the pretrained model (i.e., feature extraction mode), only training the top classifier on the target task.

We thus indicate several multistage procedures: e.g., XLM$\rightarrow \pi \rightarrow \varphi$ indicates that the XLM model is further pretrained on the in-domain tweet corpus, followed by finetuning on the end task.

4 Experimental Setup

We focus our experiments on text classification problems that are commonly dealt with by customer service teams. First, we describe our Twitter conversation corpus used for in-domain finetuning (Section 4.1), then we introduce the target tasks and corresponding datasets (Section 4.2). For most target tasks, we hold out 10% of the data for testing, while the remaining part is used for training. We then utilize 10-fold cross-validation on the training data to select optimal hyper-parameters for each end task. When the dataset comes with a predefined train-test split, we keep that. For the pretrained transformer models (RoBERTa, XLM, BERTweet), we use the publicly available 'base' versions.

4.1 Twitter Conversation Corpus

Our corpus for in-domain pretraining was crawled using Twitter's API.[2] The collected dataset is small compared to the original language models' data, but still larger than most corpora which SMEs have at their disposal. As such, it represents an easily collectable customer service dataset that SMEs can leverage to boost models on their own data. The tweets were gathered between May and October 2020. We started by gathering a list of 104 companies, all active on Twitter, in the sectors of telecommunication, public transportation, and airlines. We aimed for four different languages (English, French, Dutch, German).

We preprocessed the data by removing conversations not covering at least one client/company interaction, or containing undefined languages. We further converted usernames and links into the special tokens @USER and @HTTP_URL, respectively,

---

[2]https://developer.twitter.com/en/docs

Language	|convs|	|tweets|
English	135.1k	406.3k
French	60.9k	212.9k
Dutch	45.6k	141.0k
German	33.0k	104.5k
All	274.6k	864.7k

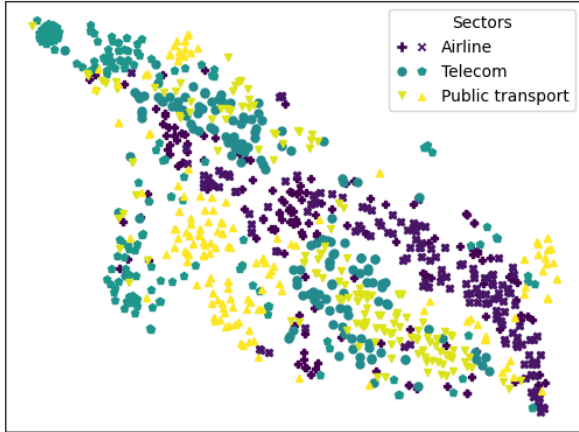Table 1: Number of collected conversations (|convs|) and tweets (|tweets|) for each language.

Figure 1: 2D visualization of the first hidden-state representations of the XLM on randomly sampled tweets for six companies, in three different sectors using t-SNE (Van der Maaten and Hinton, 2008). Each color stands for a different sector, while the marker types represent the different companies, and each data point corresponds to a single operator tweet (covering multiple languages).

and converted emojis into corresponding strings.[3] The resulting corpus contains 865k tweets over 275k conversations in the four target languages (see Table 1).

Even though our corpus contains data from different sectors, we noticed that the dialogue flow, customer intents, and structure of conversations are fairly comparable across the target sectors (cf. Fig. 1). Examples of often recurring types of tweets are expressions of gratitude towards customers, requests for information, or typical ways to reply to complaints. Hence, we expect this corpus to be useful not only for companies that fall under one of the included sectors, but also for other companies that provide customer services over tweets.

4.2 CRM Tasks and Datasets

Complaint Prediction – Timely complaint detection is of utmost importance to organizations, as

it can improve their relationship with customers and prevent customer churns. Preoţiuc-Pietro et al. (2019) and Greenleaf et al. (2015) proposed two datasets for identifying complaints on social media which contain 3,499 and 5,143 instances, respectively. The former (**Complaint-2**) covers two types of companies (airline companies and telecommunication), while the latter (**Complaint-9**) consists of data from nine domains such as food, car, software, etc. Both datasets are in English. To experiment with cross-lingual tuning for complaint prediction, we use the French complaint dataset for railway companies from (**Complaint-R**; Ruytenbeek et al. 2020). Since all their 201 conversations are labeled as complaints, for training, we complemented them with negative sampling from French railway conversations in our own Twitter corpus. For testing, we annotated 200 held-out conversations.

Churn prediction – Customer churn implies that a customer stops using a company's service, negatively impacting its growth. Churn prediction is cast as a binary classification task (churn or non-churn) on any input text. We utilize the data provided by Amiri and Daume III (2015) with tweets from three telecommunication brands, resulting in a corpus of 4,339 labelled English tweets.

Subjectivity Prediction – Detecting subjectivity in conversations is a key task for companies to efficiently address negative customer feelings or reward loyal customers. It may also serve as a filtering task for more fine-grained tasks such as emotion identification. We annotated 8,174 Dutch conversations from our Twitter corpus (Section 4.1). A dialogue is judged "subjective" if at least one of the customer turns contains emotions (explicit or implicit), and otherwise "objective".

Relevance Prediction – The goal of this task is to determine whether an incoming text is relevant for further processing or not. We use data from GermEval 2017 (Task A) which contains over 28k short length messages from various social media and web sources on the German public train operator Deutsche Bahn (Wojatzki et al., 2017). For this dataset, the evaluation is measured on two evaluation sets: one collected from the same time period as the training and development set (viz. synchronic), and another one containing data from a later time period (viz. diachronic).

Polarity Prediction – For this task, a system has to classify the sentiment that resides in a given text fragment according to polarity (positive, negative,

---

[3]We used `https://github.com/carpedm20/emoji` to convert emojis

Model	Complaint-2 (English)		Complaint-R (French)		Churn (English)		Subjectivity (Dutch)		Relevance (German)		Polarity (German)	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	$F1_{syn.}$	$F1_{dia.}$	$F1_{syn.}$	$F1_{dia.}$
Majority-class	72.4	42.0	56.0	35.8	78.4	43.9	55.0	35.5	81.6	83.9	65.6	67.2
LR (tf-idf)	83.5	77.0	57.5	57.4	85.1	71.7	71.6	70.9	88.4	87.7	71.1	70.4
SVM (tf-idf)	84.4	80.2	59.0	58.8	87.3	80.1	71.7	71.0	90.4	88.8	74.8	72.8
Reference	82.0[1]	62.7[1]	-	-	-	78.3[2]	-	-	85.2[3]	86.8[3]	66.7[3]	69.4[3]
BERTweet → ∅	80.5	71.6	-	-	79.3	55.2	-	-	-	-	-	-
BERTweet → φ	**90.0**	**86.1**	-	-	**93.0**	**90.0**	-	-	-	-	-	-
RoBERTa → ∅	77.9	74.5	-	-	78.3	59.7	-	-	-	-	-	-
RoBERTa → φ	87.5	85.1	-	-	88.4	84.8	-	-	-	-	-	-
XLM → ∅	76.2	61.6	44.0	30.5	61.6	55.7	63.8	62.4	83.1	84.7	64.5	66.8
XLM → φ	85.4	83.4	54.0	46.2	84.1	75.3	73.4	72.9	91.6	91.7	76.1	73.5
XLM → π → ∅	81.8	76.8	56.5	54.1	79.7	66.0	71.6	71.1	84.4	85.3	65.1	68.0
XLM → π → φ	86.9	82.7	**62.0**	**61.9**	87.8	83.7	**74.6**	**74.2**	**92.7**	**92.5**	**78.7**	**76.1**

Table 2: Classification results (accuracy ACC and F1-score) on CRM tasks using pretrained language models with two settings for pretraining: Feature extraction (→ ∅) and finetuning (→ φ). Missing values ('-') are due to unavailable reference scores, or a language mismatch between model and task.
[1] Greenleaf et al. (2015), [2] Amiri and Daume III (2015), [3] Wojatzki et al. (2017)

or neutral). Polarity prediction has often been applied on reviews, by predicting the attitude or sentiment of a reviewer with respect to some topic. We use the GermEval 2017 (Task B) dataset (Wojatzki et al., 2017) (cf. supra) to analyze the polarity of the Deutsche Bahn customers' feedback. We also use the polarity dataset from Sanders (2011) (**Sanders**).

5 Results and Discussion

We now present our findings for two finetuning scenarios: transfer across languages and across tasks. Section 5.1 investigates the effect of unsupervised multilingual pretraining. Section 5.2 then explores how to further improve by finetuning the pretrained language models on similar tasks.

5.1 Language Transferability

We compare the pretrained transformer experiments with the following baselines: majority-class (to get an idea of class imbalance), logistic regression (LR) and support vector machine (SVM) with tf-idf features. For the three transformer models, we compare the feature extraction setting (→ ∅) with finetuning (→ φ) on the target task. On the multilingual XLM, we measure the impact of first pretraining (→ π) on our multilingual tweet corpus, after which both transfer settings are again tested on the target tasks. Table 2 reports the results (in terms of accuracy and F1 scores), including scores from literature when available ('Reference'). It

should be noted that the reference scores are not state-of-the-art, but they are the scores communicated in the original dataset papers.

Only for the English tasks (Complaint-2 and Churn), results for BERTweet and RoBERTa are reported. The monolingual tweet-based model BERTweet outperforms all other models when finetuned on these tasks. Although a large domain-specific mono-lingual language model seems a fine choice, it may not be available for other languages. We therefore investigate the impact of a multilingual generic model (XLM was not specifically pretrained on tweets), and the impact of additional finetuning on our dedicated twitter corpus.

In general, transformer models finetuned on the end task strongly outperform frozen ones. For the non-English tasks, the model XLM→ ∅ with the frozen XLM encoder shows weak performance, in some cases below the baselines. The model XLM→ φ finetuned on the end task performs better. For the non-English tasks, the XLM model pretrained on our Twitter corpus and finetuned on the tasks (XLM→ π→ φ) in all cases outperforms the finetuned XLM by a few percentage points and the baselines by an even larger margin. The performance differences between XLM→ φ and XLM→ π→ φ clearly underscore the importance of in-domain multilingual pretraining. Furthermore, the results of XLM→ π→ φ for the English tasks suggest that additional pretraining on a moderately small, in-domain dataset can make the performance

of the multilingual XLM model comparable to the monolingual RoBERTa.

Another promising observation is that the hyper-tuned classical baselines, such as SVM, are strong competitors compared to *frozen* language models, especially on tasks that are highly sensitive to domain-specific features. For instance, for churn prediction, keywords such as 'switch to', 'quit' and 'change provider' can easily be triggered by the SVM, while *frozen* pretrained models have not learned to identify these features. This finding might be helpful to achieve better insight into the operational aspects of *frozen* neural models compared to simple classical approaches.

As a side result (not explicitly included in this work) we found that the multistage pretraining (XLM$\to \pi$) leads to better performance when incorporating multiple languages compared to a single language. The performance drops especially when training data from a single language (e.g., Dutch) is fed into the model, which is then evaluated on other languages (e.g., English).

5.2 Task Transferability

We now investigate to what extent representations tuned on a related task can help for a given target task. In particular, Complaint-9 is the end task, and we compare the effect of finetuning on the end task only, vis-à-vis first finetuning on a related task and then on the end task. For the related task, we experiment with Complaint-2 and Sanders, as shown in Table 3. We observe that there seems to be no clear merit in the additional finetuning step on a small related end task. Pretraining on our larger Twitter corpus, however, still increases effectiveness.

Test Dataset	Complaint-9		
Train Dataset	**C-9**	**C-2 & C-9**	**S & C-9**
Majority-class	39.1	-	-
SVM (tf-idf)	78.6	-	-
Preotiuc-Pietro et al.	79.0	-	-
XLM $\to \varphi$	78.6	79.3	80.1
XLM $\to \pi \to \varphi$	82.4	80.0	82.8

Table 3: F1 test scores on Complaint-9 for finetuning on the end task alone (indicated as **C-9**) vs. on Complaint-2 or Sanders and again on Complaint-9 (columns **C-2 & C-9**, respectively, **S & C-9**.).

6 Conclusion

We investigated multilingual and across-task transfer learning for customer support tasks, based on transformer-based language models. We confirmed prior insights that finetuning the models on low-resource end tasks is important. Additional pre-training on a moderately sized in-domain corpus, however, provides a complementary increase in effectiveness, especially in the non-English setting and starting from a generic multilingual language model. We provide a newly collected multilingual in-domain corpus for customer service tasks and derive the aforementioned findings from experiments using it on five different tasks.

7 Acknowledgments

References

Hadi Amiri and Hal Daume III. 2015. Target-dependent churn classification in microblogs. In *Proceedings of AAAI*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettle-moyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Lei Cui, Shaohan Huang, Furu Wei, Chuanqi Tan, Chaoqun Duan, and Ming Zhou. 2017. SuperA-gent: A customer service chatbot for e-commerce websites. In *Proceedings of ACL*.

Xin Deng, Ross Smith, and Genevieve Quintin. 2020. Semi-supervised learning approach to discover enterprise user insights from feedback and support. *arXiv preprint arXiv:2007.09303*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Dylan Greenleaf, Jianzhe Liu, Yuanbo Wang, and Antoine Rigaut. 2015. Final project report: Identifying customer complaints on social media.

Narendra Gupta, Mazin Gilbert, and Giuseppe Di Fabbrizio. 2010. Emotion detection in email customer care. In *Proceedings of WS*.

Narendra K Gupta. 2011. Extracting descriptions of problems with product and services from twitter data. In *Proceedings of SWSM*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. In *Proceedings of ACL*.

Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2019. Machine reading comprehension for answer re-ranking in customer support chatbots. *Information*.

Siu Cheung Hui and G Jha. 2000. Data mining for customer service support. *Information & Management*.

Kaliyaperumal Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2019. Cross-lingual ability of multilingual bert: An empirical study. In *Proceedings of ICLR*.

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*.

Piero Molino, Huaixiu Zheng, and Yi-Chia Wang. 2018. Cota: improving the speed and accuracy of customer support through ranking and deep networks. In *Proceedings of ACM-SIGKDD*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: a pre-trained language model for English Tweets. In *Proceedings of EMNLP*.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of ACL*.

Daniel Preoţiuc-Pietro, Mihaela Găman, and Nikolaos Aletras. 2019. Automatically identifying complaints in social media. In *Proceedings of ACL*.

Daniel Preotiuc-Pietro, Mihaela Gaman, and Nikolaos Aletras. 2019. Automatically identifying complaints in social media. In *Proceedings of ACL*.

Sebastian Ruder. 2019. *Neural Transfer Learning for Natural Language Processing*. Ph.D. thesis, National University of Ireland, Galway.

Nicolas Ruytenbeek, Ilse Depraetere, and Sofie Decock. 2020. Linguistic (in)directness in twitter complaints: a contrastive analysis of railway complaint interactions. *Journal of pragmatics*.

Niek J Sanders. 2011. Sanders-twitter sentiment corpus. *Sanders Analytics LLC*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*.

Michael Wojatzki, Eugen Ruppert, Sarah Holschneider, Torsten Zesch, and Chris Biemann. 2017. GermEval 2017: shared task on aspect-based sentiment in social media customer feedback. In *Proceedings of GermEval*.

Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Emerging cross-lingual structure in pretrained language models. In *Proceedings of ACL*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of ICCV*.