Deep Learning Models for Customer Complaint Classification

Chau-Minh Nguyen
Department of Marketing
HEC Montreal
chau-minh.nguyen@hec.ca

Mehdi Ghaniabadi
Dept. of Logistics and Operations Management
HEC Montreal
mehdi.ghaniabadi@hec.ca

Abstract

In this project, we implement novel contextualized language models such as BERT, RoBERTa and XLNet, along with other older models such as AWD-LSTM (contexualized) and biLSTM (with static word embeddings of GloVe) on a Customer Complaint Classification task. Overall, we implement 8 deep learning language models and show that they significantly outperform the other models used in the literature for this task and dataset. Our 8 models and their results are also compared in details.

1 Introduction

This project is a continuation of the research project of Master student William Blais, Professor Yany Grégoire and Associate Professor Marcelo Nepomuceno. They used different methods to classify three kinds of online customer complaints. In this project, we hope to use the state-of-the-arts language methods to increase the accuracy achieved by them.

The complaints were collected from the websites Bbb.org, Consumeraffairs.com, Ripoffreport.com, Sitejabber.com, YELP.com, Trustpilot.com and Amazon.com. The total dataset includes 4,827 complaints with 2,141 complaints of class 0 (Reparation); 1,690 of class 1 (Vigilante) and 996 complaints of class 2 (Disillusioned).

Two human coders classified the dataset. Their classification accuracy was 83.24%. Different methods have been implemented before our project. The lexicon-based method LIWK gave the accuracy of 55-56% and 93 indicators. Random Forest approach using LIWK indicators as variables resulted in 52.06% accuracy. BiLSTM stand-alone gave a 59.74% accuracy. In this project we significantly improve these results by implementing novel transformer-based language models and achieve a best accuracy of 74% by RoBERTa language model.

The rest of this report is presented as follows. In the next section we present an extensive review of the literature. In section 3, the methodology and the implemented language models of this project are introduced and compared. Section 4 presents the results of the models on our task and their analysis. The concluding remarks are given in section 5.

2 Literature review

2.1 Three groups of complaints

Looking into customers' revenge, Grégoire et al. (2009) found that online complainers' desire for revenge decreased over time, but their wish for avoidance increased over time and it was the best customers who held the longest negative reactions. Obeidat et al. (2018) found four types of online revengers: materialistic, ego-defending, aggressive and rebellious. The motivation for revenge could be the customers' perception of the firm's greed and the revenge could be direct or indirect (Grégoire et al., 2019). These customers are referred to here as vigilantes.

Using the web to tell the public about their dissatisfaction, customers tried to convince the public to join their cause by showing failures as betrayals of rights, emphasizing the seriousness of the failures, depicting firm leaders as evil doers, pointing to other customers' complaints of the firm to justify the blame, seeing themselves as crusaders for a cause and encourage other unhappy customers to be united as a group (Ward & Ostrom, 2006).

For service customers in particular, service termination would have a stronger impact on revenge among previously happy customers, while for previously unhappy customers, service demotion would have a stronger influence on revenge.

However, not all online complainants seek revenge. Weizl (2019) distinguished "constructive" complainants from "vindictive"/vigilante ones. Constructive complainants sought a response from the company and thus, were receptive to the company's reparation effort, while the vigilante ones see recovery efforts as interfering in the consumer-to-consumer conversations.

The constructive complainants are referred to here as reparation complainants. They saw a problem as a task, a mistake to be fixed, and gave priority to a solution. Their language was more formal than the vigilante customers, used "I" and the past tense more often, while the vigilante were more likely to use "you" and they" and used more often the present time (Grégoire et al., 2019).

Blais (2018), in his Marketing Master thesis at HEC Montreal, added to these two groups a third group, called the disillusioned. These complainants indicated a desire to withdraw from all interaction with the firm (McCullough et al., 1998). Their reaction was avoidance (Grégoire, Tripp et Legoux, 2009). In their effort to lessen the anxiety and frustration associated with the events, they avoided any relationship with the firm (Sundaram, Mitra et Webster, 1998). Blais, using LIWK's variables and ANOVA, found this group of complainants tended to use more words indicating time than the other two groups.

Moreover, the number of variables showing a significant difference between the vigilante and the reparation groups were greater than those indicating a significant difference between the vigilant and the disillusioned and between the reparation and the disillusioned.

Among the three groups, the reparation tended to take upon themselves part of the blame for the service failures and indicated less certainty than the other groups. The vigilante, on the other hand, were more likely to use exclamation words, more emotional and less rational. They were more likely to illustrate the company as the opposition of consumers and used impersonal pronouns. The desire to revenge is in the present, so they referred less to the events in the past. The disillusioned used fewer words of negation. They took a factual approach, regularly referred to what they went through in the past and used often the first singular pronoun "I". They did not use many persuasion words, as their complaints did not aim to persuade the firm or other consumers (Blais, 2018).

2.2 Lexicon-based classifying tools

Lexicon is the vocabulary of a person or language. Lexicon-based sentiment analysis methods use lexicons to calculate polarity of individual words and from these scores, determine the overall polarity of text. Lexicon-based approaches could be divided into two sub-approaches: corpus-based and dictionary-based. The corpus-based approach is context-oriented rather than word-oriented. The dictionary-based approach combines current sentiment words in the dictionary and newly found words to add to the list. (Yadav & Pandya, 2017).

Hartman et al. (2018), when comparing automated text classifying methods, proved that lexicon-based methods (LIWK, VADER, AFINN, BING and NRC) lagged behind machine learning methods (ANN, kNN, Naïve Bayes, Random Forest and SVM) in sentiment analysis.

2.3 Linear textual classification

In Naïve Bayes, one approach is to choose the weights to maximize the joint probability of a

training set of labeled documents. This is called maximum likelihood estimation. We also assume that the dataset is independent and identically distributed (Eisenstein, 2018).

Another linear classification approach is support vector machines. According to Joachims (1997), this method suits text categorization, due to its ability to deal with certain characteristics of text: high dimensional feature spaces, few irrelevant features and sparse document vectors.

While Naïve Bayes is probabilistic and support vector machine are discriminative, logistic regression combines the power of probabilistic and discriminative classifiers.

2.4 Deep Learning in textual classification

In a convolution neural network (CNN), convolution layers, also called kernels, go through the word embedding matrix. Each kernel extracts a certain pattern of n-gram. Then, max-pooling sub-samples the input data, by running a max-operation for each layer. On one hand, max-pooling always maps the input to fixed dimensions of the outputs. On the other hand, it reduces the output's dimensionality, while still maintaining the most important n-gram features. Word embeddings could be randomly initialized or pre-trained on a large corpora. A convolution layers combined with max-pooling are stacked to make a deep CNN. This approach allows for sentence representation (Young et al, 2018).

A recurrent neural network (RNN) with "its sequential processing by modeling units in sequence," is suitable for the sequential characteristics and context dependencies of language (Young et al., 2018). RNN allows for modeling whole sentences, with the main contents assigned to a set dimension hyperspace. RNN could also support "time distributed joint process" (Young et al, 2018). The related tasks tend to be part-of-speech tagging and could range from multi-label tag classification to sentiment analysis. (Santos & Zadrozny, 2014; Chet et al., 2017; Poria et al., 2017; Tong et al., 2017).

For RNN, the most important element is often its hidden state. It is the network's memory element and gathers information from other time-steps. As a simple RNN structure often has to deal with the vanishing gradient problem, other RNN architectures such as long-short term memory (LSTM), gated recurrent units (GRU) and residual networks (ResNets) address this shortcoming. LSTM and GRU, in particular, are widely used. (Young et al., 2018).

A limitation of the conventional encoder-decoder framework is that there are occasions when the encoder encodes information that might be unnecessary for the task. This gave rise to the attention mechanism. It lets the decoder attend to the input at each step of generating the output. (Young et al., 2018).

Considered a recent breakthrough in NLP, Bidirectional Encoder Representations from Transformers (BERT) was introduced and then open-sourced by Google (Devlin et al., 2018). Its main contribution was using the bidirectional training of the Transformer, thanks to a technique called Masked LM. Instead of reading from left to right or right to left as in directional models, BERT encoder reads the entire text sequence (Horev, 2018).

2.5 Customer complaint classification

Research on classifying consumer complaints has been quite limited. HaCohen-Kerner et al. (2019) used Bayes Network, SimpleLogistic, SMO and Random Forest to classify complaint letters in Hebrew according to seven company categories. Thomas (2018) used LSTM to classify US Consumer Finance Complaints into related subjects such as mortgages, student loans and credit reports. Classifying complaints on construction quality problems according to the relevant government department, Zhong et al (2019) found that Convolution Neural Network had higher average precision, recall and F1 scores than the Bayes-based and SVM classifiers.

In other words, so far, complaint classifying in literature has been limited to categorizing according to objective criteria. In our case, the criteria are subjective, as they required the coders/algorithm to predict the motivations of the complainants. This explained why the agreement between two human coders was only 83.24%.

Furthermore, the classes are not balanced, with the minority class, the disillusioned accounting for only 20.6% of the whole dataset. Earlier findings also showed that there were fewer indicators distinguishing this group from the other two groups. In other words, it might be more difficult to classify this class than the other two classes.

3 Methodology and models

In classical natural language processing (NLP) models, static word embedding models such as GloVe or Word2Vec are used to train word vectors. Then, such word embeddings are used in classical models like RNN or LSTM in order to perform NLP tasks such as text classification and sentiment analysis. The main shortcoming of such word embedding models is that they are static and give the same vectors for the same word irrespective of its context; however, a single word can have a totally different meaning depending on its context. As an example, consider the sentence: “The doctor arrived at the prison cell with his cell phone in order to take blood cell samples from prisoners”. In this sentence the word “cell” has three different meanings depending on the context, nevertheless static word embedding models provide a single vector for the word “cell”.

To diminish the above shortcoming, contextualized models such as Elmo and BERT are introduced which generate different vectors for a word like “cell” depending on its context. They can also be optimized dynamically for specific tasks. In fact, currently BERT is the most well-known contextualized NLP model which is pretrained on a large corpus and has been shown to generate high quality word embeddings and also achieves state-of-the-art performance on various NLP tasks, although it can be quite computationally expensive.

Since the introduction of BERT in October 2018, various researchers have tried to improve its performance in regards to either computational time or prediction accuracy by introducing new contextualized and transformer-based language models, among which, DistilBERT, XLNet and RoBERTa can be considered as the most notable ones. In particular, XLNet and RoBERTa try to provide a better prediction accuracy while they may require 4 or 5 times more training time compared to BERT. DistilBERT on the other hand improves in regards to computational speed of BERT, while it may be slightly less effective in terms of prediction accuracy. Therefore, currently the literature lacks a language model which performs better than BERT in both accuracy and computational time. One of the main reasons behind the better accuracy performance of XLNet and RoBERTa is that they both use a larger training corpus, 113GB and 160GB, respectively, while BERT and RoBERTa utilize a 16GB corpus size. We will show that in our project, which is a sentiment analysis task on a new dataset, RoBERTa provides the best test accuracy.

In addition to the four language models mentioned above (BERT, DistilBERT, XLNet and RoBERTa), we also use four other older models in order to make sure a variety of deep learning models are used to find the best model for the sentiment analysis task on our dataset, and also to compare the results of different types of NLP models and examine whether the novel models such as BERT or XLNet in fact outperform the classical models. More specifically, we use AWD-LSTM which is a contextualized language model and is introduced in August 2017 which was one year before BERT. Then we use biLSTM which is a static language model and uses GloVe word embeddings. The GloVe model for word embeddings is introduced in 2014. Moreover, Naïve Bayes SVM (NBSVM), introduced in 2012, is also implemented which uses a totally different type of text representation called Term Frequency-Inverse Document Frequency (TF-IDF vectors). Finally, an ensemble of the two last models are also considered (biLSTM+ NBSVM).

Our dataset contains roughly 4900 number of rows, and 3 number of columns which correspond to ids, the complaints, and labels for each complaint. We divide our dataset into a train set with 90% of the original dataset, and a test set with 10% of the original dataset. 10% of the train set is also used as the validation set in order to tune the hyper-parameters (such as learning rate, number of epochs, maximum length, etc.) of each language model based on the validation accuracy. Then, the test accuracy of each tuned model is also calculated which is the main benchmark to evaluate the performance of our NLP models. The results on the test set for each model is presented and analyzed in the next section.

It is worth mentioning that we also used LDA (Latent Dirichlet allocation) to find 9 main interpretable themes of the complaints. From the top twenty words for the topics, the following themes could be interpreted: games (1), loans and mortgage (2), good quality (3), groupon (4), subscription (5), electronic products (6), pets (7), time (8), and foods (9).

For the two methods Roberta and XLNet, we used the library Fastai, which provides user-friendly architecture for transformer methods. Each model included 14 layer groups: one for embedding, 12 for transformers and one for classifying. We ran three cycles, and 1 to 2 epochs for each cycle. Liu et al. (2019) developed RoBERTa, which built on the masking strategy of BERT (Devlin et al. 2018). RoBERTa changed important hyperparameters of BERT and resulted in better task performance. XLNet, offered by Yang et al. (2019), also aimed to overcome BERT's limitation of pre-trained and finetuning discrepancy. For DistilBERT (Sanh et al. 2019) and AWD-LSTM (Merity et al. 2017), the library Fastai is also used, and the learning rate and number of epochs are optimized for our task.

Fastai allows for mock training to find a suitable initial learning rate. Often the suggested learning rate was not as effective as what we finally chose, such as the learning rate of 1e-3 or 1e-4. For the training, we used different learning rates for different cycles. We also used freezing for certain layers during training to prevent well-trained layers from being modified.

NB-SVM is another novel solution by combining Naïve Bayes (NB) with Support Vector Machines (SVM), developed by Wang and Manning (2012). Both are baseline methods for classifying text, with NB having an edge in short text and SVM in longer text.

Another method we used is combining GloVe with BiLSTM. By using GloVe vocabulary as embedding, we created context for the LSTM method, thus improved the method accuracy.

And finally, we combined GloVe BiLSTM with NB-SVM, by using the average of the probability of each method predicting complaints to belong to a class. By combining a deep learning method with a linear one, we hope to take advantage of the strength of each approach.

4 Results

The table below shows the results of our 8 NLP models on the test set according to various classification metrics.

Table 1: The classification results of the implemented deep learning models

	BERT	**Distilled BERT**	**RoBERTa**	**XLNet**	**AWD-LSTM**	**NBSVM**	**GloVe +BiLSTM**	**NBSVM +GloVe +BiLSTM**
Precision								
Average	0.72	0.72	0.74	0.73	0.68	0.67	0.66	0.66
Reparation (0)	0.73	0.73	0.76	0.75	0.68	0.61	0.7	0.67
Vigilante (1)	0.75	0.72	0.78	0.74	0.73	0.69	0.66	0.68
Disillusioned (2)	0.66	0.70	0.65	0.67	0.62	0.75	0.56	0.62
Recall								
Average	0.72	0.72	0.74	0.73	0.69	0.64	0.66	0.66
Reparation (0)	0.85	0.81	0.77	0.77	0.78	0.87	0.68	0.75
Vigilante (1)	0.72	0.73	0.82	0.77	0.75	0.65	0.78	0.78
Disillusioned (2)	0.50	0.54	0.59	0.61	0.43	0.21	0.46	0.35
F1 Score								
Average	0.72	0.72	0.74	0.73	0.68	0.61	0.65	0.65
Reparation (0)	0.79	0.77	0.76	0.76	0.73	0.71	0.69	0.71
Vigilante (1)	0.74	0.72	0.8	0.75	0.74	0.67	0.71	0.72
Disillusioned (2)	0.57	0.61	0.62	0.64	0.51	0.33	0.51	0.45
Accuracy	**0.725**	**0.72**	**0.74**	**0.73**	**0.68**	**0.64**	**0.66**	**0.66**

The results showed that RoBERTa did best in terms of overall accuracy (74%), while NB-SVM did worst (64%). Combining NB-SVM with BiLSTM did not help significantly with the accuracy, as the overall accuracy was similar to that of Bi-LSTM (66%). AWD-LSTM which is a contextualized model, performed better than both biLSTM and NBSVM, but was outperformed by newer dynamic language models like XLNet and BERT.

As predicted, all the methods had difficulty predicting the class "Disillusioned", which is apparent in all our metrics presented in the table. This is probably because it is both a minority class and one with fewer features to distinguish it from the other two classes. Over all, this class has a lower score in all three classification metrics of Precision, F1 score, and Recall. The only exception is NB-SVM, with a higher precision score for the "Disillusioned" than the "Vigilante" and the "Reparation".

Moreover, DistilBERT was more computationally efficient than BERT, as it is supposed to, but its accuracy was slightly lower than BERT (by 0.5%). Both XLNet and RoBERTa outperformed BERT in terms of accuracy.

We also noticed that the famous model BERT has a restriction on the length of its inputs which is 512. Some of the complaints in fact had a length more than 512. Nevertheless, when we examined different values for this hyperparameter (max_length), we realized that higher max_length does not give better accuracy after a certain point which was a maximum length of around 350 in our case. It may seem unusual and even counter-intuitive since typically in machine learning models more data helps. However, in our case, it seems to be due to the fact that the early phrases give the most relevant information about the sentiment of the user. So, it is interesting to note that more data does not necessarily help the NLP models. In fact, the quality and relevance of the added data is also important.

Overall, our results show that such new NLP models like BERT and RoBERTa outperform the classical language models significantly, but still require improvements on difficult NLP tasks which include language subtleties like our project data, in order to achieve a result close to human judgment.

5 Conclusions

Considering the difficulty of the classification problem, our methods did reasonably well. Even the simple linear method NB-SVM scored a better result than the previously tested methods.

Considering the fact that there is still very limited research in customer complaint classifications and none so far for classification of subjective complaint classes, our work has the potential to contribute new understandings of using machine learning general, and deep learning, in particular, in marketing research.

However, there is still a gap between the accuracy of our methods (74%) with the human accuracy (83.24%). We were not able to make use of some already available data. In future research, such variables as the LIWK indicators, the different forums to which the complaints belonged and the topics identified by LDA could be useful in other state-of-the-arts approaches, for example using graphs to classify text.

References

Blais, W. (2018). Une analyse textuelle des principaux types de plaignants en ligne, une comparaison entre les justiciers, les conciliateurs et les disabuses.

Chen, G. et al. (2017). Ensemble application of convolutional and recurrent neural networks for multi-label text categorization," in *Proc. Int. Joint Conf. Neural Networks*, 2017, pp. 2377–2383.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Obeidata, Z.M. et al. (2018). Social media revenge: A typology of online consumer revenge. *Journal of Retaling and Consumer Services*.

Grégoire, Y. et al. (2018). What Do Online Complainers Want? An Examination of the Justice Motivations and the Moral Implications of Vigilante and Reparation Schemas. *Journal of Business Ethics.*

Grégoire, Y., Laufer, D. & Tripp, T. (2010). A comprehensive model of customer direct and indirect revenge: understanding the effects of perceived greed and customer power. *Journal of the Academy of Marketing Science.*

Grégoire, Y., Tripp, T.M. & Legoux, R. When Customer Love Turns into Lasting Hate: The Effects of Relationship Strength and Time on Customer Revenge and Avoidance. *Journal of Marketing.* Vol. 73, No. 6 (Nov., 2009), pp. 18-32

HaCohen-Kerner, Y. et al. (2019). Automatic classification of complaint letters according to service provider categories. *Information Processing and Management*, 11/2019, Volume 56, Numéro 6

Haenel1, C.M. et al. (2019). The Perils of Service Contract Divestment: When and Why Customers Seek Revenge and How It Can Be Attenuated. *Journal of Service Research 2019*, Vol. 22(3) 301-322

Hartman, J., Huppertz, J., Schamp, C. & Heitmann, M. (2019). Comparing text classification methods. *International Journal of Research in Marketing*, vol. 36, no 1, p. 20-38

Joachims, T. (1997). Text Categorization with Support Vector Machines: Learning with Relevant Features. *Research Reports of the unit no. VIII* (AI. Computer Science Department. University of Dortmund

Liu, Y. et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692v1*

McCullough, E. Michael., Rachal, K. Chris., Sandage, J. Steven., Worthington Jr. L. Everett., Wade Brown, Susan., Hight, L. Terry. (1998). Interpersonal Forgiving in Close Relationship: Vol. 2: Theoretical Elaboration and Measurement, *Journal of Personality and Social Psychology*, vol. 75, p. 1586-1603.

Merity, S., Keskar, N. S., & Socher, R. (2017). Regularizing and optimizing LSTM language models. *arXiv preprint arXiv:1708.02182*.

Poria, S. et al. (2017). Context-dependent sentiment analysis in user-generated videos," in *Proc. Annu. Meeting Association Computational Linguistics*, pp. 873–883.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Santos, C.D. & Zadrozny, B. Learning character level representations for part-of-speech tagging," in *Proc. 31st Int. Conf. Machine Learning*, 2014, pp. 1818–1826.

Sida Wang and Christopher D. Manning. (Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. *ACL 2012*.

Sundaram, D. S., Mitra, K., Webster, C. (1998). Word-Of-Mouth Communications: a Motivational Analysis, *Advances in Consumer Research*, vol. 25, p. 527-531.

Thomas, N.T. (2018). A LSTM based Tool for Consumer Complaint Classification. *International Conference on Advances in Computing, Communications and Informatics (ICACCI).*

Tong, E. et al. (2017). Combating human trafficking with deep multimodal models. *ArXiv Preprint, arXiv:1705.02735*.

Ward, C. James., Ostrom, L. Amy. (2006). Complaining to the masses: The role of protest framing in customer-created complaint web sites. *Journal of Consumer Research, vol. 33, no 2, p. 220-230*.

Weizl, W.J. Webcare's effect on constructive and vindictive complainants. *Journal of Product & Brand Management*, 05/2019, Volume 28, Numéro 3

Yadav, P. & Pandya, D. (2017). SentiReview: Sentiment Analysis based on Text and

339 Emoticons. *International Conference on Innovative Mechanisms for Industry Applications*.

340 Yang, Z. et al. (2019). XLNet: Generalized Autoregressive Pretraining for Language
341 Understanding. *arXiv:1906.08237v2*

342 Young, S. et al. (2010). The hidden information state model: A practical framework for
343 POMDP- based spoken dialogue management. *Comput. Speech Lang*., vol. 24, no. 2, pp.
344 150–174, June 2010.

345 Zhong, B. et al. (2019). Convolutional neural network: Deep learning-based classification of
346 building quality problems. *Advanced Engineering Informatics*. Volume 40, April 2019,
347 Pages 46-57