Empowering Borrowers in their Choice of Lenders: Decoding Service Quality from Customer Complaints

Aniruddha M. Godbole
agodbole@indiana.edu
School of Informatics, Computing, and Engineering
Indiana University
Bloomington, Indiana

David J. Crandall
djcran@indiana.edu
School of Informatics, Computing, and Engineering
Indiana University
Bloomington, Indiana

ABSTRACT

When shopping for lenders, most consumers choose a financial institution based on just a few key factors: the interest rate, the distance to the lender's nearest branch, an existing relationship with the lender, and the reputation of that lender. But most consumers fail to consider an important element that will be key to their long-term satisfaction: whether the customer service provided by the lender is commensurate with the price. Our underlying assumption in this paper is that a consumer's personality traits are associated with the issues they will face. We use state-of-the-art cross-domain word vector space mapping and representative trait vectors in this space to estimate ten personality traits corresponding to each text and use topic modeling for finding the topics in a complaint. We then use two modified collaborative topic regression methods to create two complaint topic trait spaces for each lender, and test our underlying assumption by using statistical tests for this unsupervised learning problem in three cases: mortgage loans, student loans, and payday loans. We propose that lenders could be recommended for a specific user by analyzing this space, recommending a lender with the fewest number of complaints per retail customer of that lender in the complaint space neighborhood of the customer. We suggest future work that may be undertaken for the three types of loans, including the possibility that lenders evaluate their service from a customer's perspective to track customer satisfaction over time, and extensions to other parts of the service economy.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; *Data analytics*; • **Social and professional topics** → User characteristics; • **Human-centered computing** → Collaborative and social computing.

KEYWORDS

Complaint, finance, customer service, personality traits, cross-domain word vector mapping, recommender system, collaborative topic regression, unsupervised learning

WebSci '19, June 30-July 3, 2019, Boston, MA, USA

ACM ISBN 978-1-4503-6202-3/19/06…$15.00
https://doi.org/10.1145/3292522.3326021

ACM Reference Format:
Aniruddha M. Godbole and David J. Crandall. 2019. Empowering Borrowers in their Choice of Lenders: Decoding Service Quality from Customer Complaints. In *11th ACM Conference on Web Science (WebSci '19), June 30-July 3, 2019, Boston, MA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3292522.3326021

1 INTRODUCTION

Historically, lenders have either not been evaluated by their customers for quality of service, or they have been assessed using aggregate-level reports based on subjective ratings in surveys. The U.S. Consumer Financial Protection Bureau [13] reports that almost half of all borrowers do not shop around when arranging a mortgage loan. In the absence of an easy way to compare lenders based on their customer service, consumers tend to compare based just on price (interest rate), the existence of a relationship with the lender, the distance to the nearest branch, and the reputation of the lender [13]. In general, whether for mortgages or other types of loans, there is almost no emphasis on customer service [6, 15, 51].

There is thus a need to create tools and information for consumers to make better-informed decisions based on the quality of service provided by a lender, and how it will conform with their own expectations. While machine learning has been applied extensively to the financial sector, we are not aware of any work that analyzes complaints from a customer's perspective. Even the customer-centric use cases of credit scoring, client-facing chatbots, and selling of insurance to customers [11] are actually formulated from a lender's perspective.

Our goal in this paper is to take a first step towards testing whether it is possible to automatically recommend lenders to a particular consumer based on which lenders are least likely to lead to consumer complaints. To do this, we apply machine learning and data mining techniques to a large-scale dataset of consumer complaints, looking for patterns (topics) in them. We also analyze a large-scale dataset of Twitter text, trying to infer personalities traits of individual users from it. We then look for connections between these personality profiles and the complaints for different lenders. We are not trying to resolve the complaints [25] better or faster. Rather, our goal is to try to identify lenders that are likely to give fewer reasons for making complaints.

1.1 Research Questions and Scope

In particular, we investigate two specific research questions:

(1) Is the association among the set of personality traits and complaint topics different for each lender?

(2) Is it feasible to make personalized recommendations to customers about suitable lenders based on customer service?

To study these questions, we use three specific datasets: (1) the Consumer Financial Protection Bureau (CFPB) Complaints Dataset (as of 31 July 2018) [2], (2) the TREC 2011 Microblog Dataset [41], and (3) the World Well-Being Project's [50] correlations between personality traits and words. We restrict our attention to three specific types of loans: mortgage loans, student loans, and payday loans. We do not consider temporal modeling aspects, nor do we benchmark the proxy personality trait scores. The impact of lenders selling their loans and borrowers refinancing their loan is considered to neutralize each other — for groups of customers and groups of lenders — as a simplifying assumption.

1.2 Contributions

Our main contributions are:

(1) We show it is feasible to use machine learning on a publicly available complaints dataset to empower retail customers with personalized recommendations about service providers based on their customer service;
(2) We find a general association between the set of personality traits and complaint topics in the case of payday loans, and show how lenders differ through visualizations of the latent space;
(3) We undertake domain adaptation within the English language using Cross-Domain Word Vector Space Mapping;
(4) We propose to recommend lenders based on analyzing the neighborhood around the customer in latent topic space;
(5) We propose to use a large number of topics that include a few that are semantically interpretable by the consumer;
(6) We compare different techniques for building a joint space between the topics and and personality traits; and
(7) We make our code publicly available at https://github.com/godboleam/service-quality.

2 RELATED WORK

So far, the focus in work related to complaint data has generally been on the lender's usage of the data [46, 49]. Financial firms have reportedly lobbied against a publicly available complaints dataset [20]. The focus has been to look at the dataset from the lenders' perspective [35] even though the intent of the CFPB has clearly been in favor of giving primacy to the interests of the borrower. An unspoken adversarial relationship between lenders and customers is unnecessary and perhaps short-sighted [10]. Given that the CFPB oversight seems to have not affected the overall volume of mortgage lending [24] there may not be a trade-off between lenders' interest and customers' interests even in the short-term.

Customers often regret — when they even exercise a choice — an interest rate-based selection of a lender [30]. J.D. Power Ratings says top-performing banks have fewer reported complaints and problems [31]. Their 2017 study was based on more than 5,784 responses. Deloitte suggests modification to products and processes based on complaint analysis [21], and this development is encouraging even though it is from a lender's perspective.

Despite our best efforts we could not find prior literature that explores the CFPB complaints dataset for either exploring the association between personality traits and complaint topics, or for making personalized recommendations to customers based on a lender's customer service.

3 METHODS

Our goal is to analyze a large-scale, publicly-available dataset of complaints against lenders, and to develop a word embedding space to connect the complaints to personality traits inferred through analyzing consumers' Twitter feeds. A major challenge in making this connection is that the style and vocabulary of complaints is typically very different from those of informal tweets. Moreover, we need complaints topics that are interpretable by humans, and we need to evaluate the proposed methodology without labeled ground truth data.

Our overall approach is as follows. We created word embeddings separately for the complaints text and the tweets text. Only the English tweets were considered. We then used a cross-domain word vector space mapping so that domain adaptation helps in two cases: for finding word vectors of words that are correlated with the five personality traits (on bipolar scales), and for finding the proxy traits for a Twitter users. We separately use topic modeling for the complaints dataset (without using the word vectors as [22] was not conclusive about using word vectors for topic modeling). We use the probability of that topic as a score for that topic in the complaint. We then use modified Collaborative Topic Regression (CTR) by two methods to build a joint space between complaint topics and personality traits.

We now describe the approach in more detail.

3.1 Definitions

We define a *complaint user* to be a retail customer who has made a complaint in our dataset. A *Twitter user* is a prospective borrower or someone who has sought a personalized recommendation in the past. We assume that the text of the tweets of a user reflect something about the person's personality. The *proxy trait scores* are the ten scores associated with the five personality traits, on bipolar scales. The *complaint language* is based on the complaint narrative texts in the CFPB complaints dataset. This includes all complaints irrespective of the type of the loan. The *Tweet language* is based on the tweet text (detected as English) in the TREC 2011 Microblog Dataset. Although complaints and tweets are written in the same language (English), the styles and vocabularies are quite different, as if they were different dialects. We build a *Complaint Topic Trait Space* to connect complaints and personality traits.

3.2 Data

We use three main datasets, collected as of July 31, 2018.

3.2.1 Complaints data. Around 1.5 million complaints [16] have been sent to the CFPB, and the publicly available dataset has 1,087,269 (1.08 million). This is probably the largest dataset of its kind. About 30% (0.307 million) have accompanying narratives or unstructured complaint text. We believe this complaints dataset can be considered a reasonable proxy for measuring customer service. It is believed

that the pain for a loss is felt much more than the joy felt for a similar magnitude of gain [32]. This asymmetrical relationship implies that complaints are very valuable for evaluating customer service, and a recent JD Power survey reached this conclusion for the financial sector in particular [31]. A variety of negative emotions are seen in almost half of the complaint narratives about lenders [23].

We use the complaints data independently to generate word embeddings and for topic modeling. In the case of the word embeddings, we apply the following pre-processing steps: (1) convert to lower case, (2) drop most punctuation and symbols (but retain ', $, and %), (3) replace all numbers with ∗, (4) replace all tokens such as XXXXX (which indicate private information that was scrubbed by the CFPB before releasing the data) with &, (5) remove extra spaces, and (6) use utf-8 encoding. We then apply the fastText [12] Python wrapper to create the word embeddings.

The complaints dataset does not include unique identifiers for each customer, so we cannot detect if a single customer makes multiple complaints. We thus make the (naive) assumption that each customer in the dataset has made exactly one complaint.

3.2.2 World Well-Being Project data. We use the gender and age-controlled list of 1-gram word correlations for the five personality traits (on bipolar scales) from the World Well-Being Project [50]. Most of these are words used in informal English, like the language often used on Twitter.

3.2.3 Tweets data. We use the TREC 2011 Microblog dataset [41], which had 10,617,146 (10.6mn) tweets as of mid-2018. The tweets corpus was downloaded using Twitter Tools [39] and the TREC 2011 Twitter Collection Downloader [5]. Language detection done using a port of Google's language detection library to Python [3] indicated that around two-thirds of the tweets are not in English. Only the English tweets were considered for creating the tweet language word embeddings, and we used the following preprocessing before applying fastText [12]: (1) drop most symbols and punctuation (but retain ', !, and @), (2) replace all URLs by ^, (3) replace all numbers by ∗, (4) replace all Twitter handles with @, (5) remove extra spaces, and (6) use utf-8 encoding. We believe that our collection of over one million tweets will probably suffice for building word embeddings, as more will likely not give significantly superior results [38].

3.3 Word Vector Space and Cross-Domain Word Vector Space

We address the challenge of inferring personality traits from complaint and Twitter text using cross-domain word vector space mapping, and by using personality trait proxies (see Section 3.5) in this same space. The fastText Python wrapper was used to create a complaint language word vector space and a tweet language word vector space, both of 200 dimensions. Earlier uses of cross-domain vector spaces has focused on unsupervised translation between two languages [9, 37]. We used the state-of-the-art Vecmap opensource project code [1] to create a mapping from the complaint language space to the tweet language space. This algorithm includes Cross-domain Similarity Local Scaling (CSLS) proposed by Lample et al. [37] We used the 'identical' parameter to specify that words in the tweet text that are common to the complaint text ought to have the same meaning. The Vecmap output word2vec [42] embeddings were consumed using the gensim [45] library.

	CFPB Classification as available in the dataset	
Our Classification	Product	Sub-product
Mortgage loan	Mortgage	Conventional home mortgage Conventional fixed mortgage Conventional adjustable mortgage (ARM)
	Debt collection	Mortgage Mortgage debt
Student loan	Student loan	Federal student loan servicing Private student loan Non-federal student loan
	Debt collection	Federal student loan Federal student loan debt Non-federal student loan Private student loan debt
Payday loan	Payday loan	–
	Debt collection	Payday loan Payday loan debt

Table 1: Mapping from CFPB loan type classifications to our three classifications (mortgage, payday, student loan).

3.4 Topic Modeling of Complaints

Out of the 1.08 million complaints in the dataset, the total number of complaints with a narrative is 307,120. Of these, 128,314 are such that the lender's response is not disputed by the customer. We used only the undisputed complaints for topic modeling, since these are more likely to be genuine complaints and thus higher-quality data. We considered three types of loans, mortgage loans, student loans, and payday loans.

The CFPB's classification and nomenclature of products and issue options underwent a change in April 2017 [14]. Also, the product and debt collection are identified separately by the CFPB. The map of our classification of the loans to the CFPB classification is given in Table 1. The nomenclature, both prior to and after April 2017, is aligned with how banks are organizationally structured, which means they are designed from the perspective of efficient issue resolution by a lender and not from a customer's perspective.

Of the over 100,000 undisputed complaints, we have 12,772 for mortgage loans, 8,687 for student loans, and 2,698 for payday loans. For student loans, we considered only lenders who are involved in both product and debt collection of both Federal and private student loans, to be able to analyze the complaints throughout the life of the product. We assume that a company is involved with a product or with debt collection if there is at least one complaint against it.

About 58% (76) of the mortgage debt collection companies are involved in both mortgage product and mortgage debt collection. There are 92 companies associated with Federal student loan debt collection, 154 with Private student loan debt collections, and 54 who do both. Thirty-six companies are associated with both Federal and private student loans and have at least one undisputed complaint with a narrative. About 60% (243) of the payday product companies are involved in both payday loans and payday loan debt collection.

bank payment loan go i from my saying download being info XXXX pay student company account paying loans payments I mortgage the home house property We The XX because Bank America Wells Fargo They ,u''n t'' $ time told \n\n As money debt he day work said called calls phone number He payday Mae Sallie month debt receive pay go send ask sell borrower facility time year month robo middle check debt daughter week year session pg xxxxpage mis friend credit yard human citi acre someone i. chairman name manner refer corporation grandmother cc cycle people girlfriend onset resource preciousdriver gal man daughter georgia agi boundrie set thank ed bill significance bless what i k fedloan hotel crew head performer c. bofa. hail agent me.all that:1.1 crystal statusdespite layer hr midst absence situation error modification loss!?!?!?!. doing grandmother retirement st husband sir girl what i lady claiming senior colleague birth use drafting sake red path navient yrs xxxxgreatlake ring thousand dog d definite minus bofa. co. speaker head difference box sewer spart loan/ init service director boss site means est brother hundred kin noon yesterday essential avenue tomorrow sent whom slrp female since brotherinlaws page usefuleness trid purple tid elbow lilac circuit sending abbreviation woman door name because weekend wk basic thing rank navient someone 48hrs dinner it me progress member dismiss

Figure 1: List of 222 custom stop words.

3.4.1 Preprocessing. For topic modeling, we used spaCy [27] to tokenize complaints. Only the lemmatized version of tokens tagged as nouns by spaCy were used, as these tokens generally seemed to offer better interpretability [22]. We then applied Hierarchical Dirichlet Process (HDP) topic modeling using the gensim library to identify 150 topics for each of the three types of loans. We found in initial work that the topics were often dominated by frequent but not distinctive words, making it difficult to interpret the topics. To refine our topic models, we used a heuristic procedure that iteratively identified these words and added them to a list of stop words:

(1) An HDP model's $150 \times k$ inference matrix, where k is the number of complaints (sampled from the complaints for that type of loan, sampling was required as spaCy—used for tokenization and Parts of Speech tagging—by default works with up to a million words) was fetched using the gensim library. We used the spaCy default stop word list in the first iteration.
(2) The probabilities in this matrix were set to zero if they were in the bottom or top quartiles (otherwise the interpretability of topics became difficult when attempted at the end of an iteration).
(3) For each topic the median probability (from among the k complaints) was considered to be representative for that topic.
(4) We then found the top 10 topics.
(5) Based on the keywords in the 10 topics (across mortgage loans, payday loans, student loans), appropriate additional stop words were added to the stop words list.
(6) Repeat above steps five times.

In all, 222 additional custom stop words were added to the default spaCy list in the four iterations. The topic modeling was undertaken five times for each of the three types of loans. The list of the custom stop words are shown in Figure 1.

3.4.2 Scoring. A complaint topic score is the probability of a specific topic being associated with a specific complaint. We did not rescale these scores as this would force all topics to have similar scores, which would implicitly make the unnecessary and unreasonable assumption that all topics are equally important.

Chang et al [19] found a trade-off between predictive perplexity and interpretability of latent topics. In our work, partial interpretability — i.e. having only a few of the 150 topics as interpretable — is acceptable because we only need a small set of topic importance scores from a Twitter user. The higher the relative importance score, the more important the issue/topic to that customer.

3.4.3 Sample topics. To give an idea of the topics found by our analysis, we give two examples. The top seven words in Topic #13 in our analysis included the words disbursement, afternon, violationno, confirmation, correspondence, bankrupty, and postcard, all with probabilities of about 0.003. This topic seems to correspond with complaints about rights under bankruptcy not being respected. Topic #141 seems to correspond with fraudulent advertisement, with words heat, fault, 839,i, work, checking, advertisement, and clerk, again with probabilities about 0.003.

3.5 Personality Trait Proxies

Big5 is a popular and standard personality test [18] that uses bipolar scales for extraversion, neuroticism, agreeableness, conscientiousness, and openness. Correlations between bipolar scales of the five personality traits and associated words are available [50]. Our intuition behind developing personality trait proxies is to infer personality trait-like information for word embeddings in the cross-domain vector space by considering the similarity of these word embeddings with the proxy representative trait vectors. For each of the ten (two polarities for each of five scales) traits, we compute a weighted average word vector in the vector space, yielding the ten proxy representative trait vectors. The weights are the correlations between the personality trait and the words associated with that personality trait. For each word for which there exists a word vector in the cross-domain word vector space, we find that word's similarity with a proxy representative trait vector. The same is done for other words in the given text (which could be a complaint narrative or a tweet). Then we take the average of the similarities to get a single proxy trait score for that text, and repeat for the other nine traits.

3.6 Modified Collaborative Topic Regression

Human interpretability of topics is useful in practice so that a prospective borrower could give her input on specific topics that are of interest to her, although it is sufficient for only a subset of the 150 topics to be interpretable. We address the challenge of human interpretability with a modification to Collaborative Topic Regression [54]. In the case of each complaint, we compute 160 scores: 150 scores corresponding to each of the 150 topics and 10 scores for each of the proxy personality traits. We use root mean squared error (RMSE) as the loss function given that we are doing a regression [55]. The complaints are mapped to a Complaint Topic Trait Space using the predicted values for the 160 scores. We use the Funk SVD implementation of the Surprise [28] library. Additionally, we use a hybrid (neural network) implementation of the Spotlight [7] library which uses a bilinear neural network — which in turn uses the Pytorch library — for another such Complaint Topic Trait Space. In all we have two Complaint Topic Trait Spaces for each of the three types of loans: mortgage loans, student loans and payday loans.

Type of loan	Sample Size	Complaint Topic Space	RMSE
Mortgage	12,772	Funk SVD	0.053
		Hybrid	0.054
Student	8,687	Funk SVD	0.055
		Hybrid	0.055
Payday	2,698	Funk SVD	0.059
		Hybrid	0.059

Table 3: Results of collaborative topic space regression, for different types of loans and different approaches.

For Funk SVD, we considered 4 latent factors, one epoch, and default values for all other hyperparameters. The hybrid method consists of a set of independent fully-connected layers for the user, and a set of independent fully-connected layers for the scores (complaint topics, personality trait proxies), and the outputs of these two sets are combined by a dot product [36]. We increased the regularization parameter (as compared to the default values of the Spotlight library) in order reduce the instability in the prediction space [17]. Key hyperparameter values for the hybrid method are given in Table 2. Table 3 reports the RMSEs for different approaches.

Table 4 presents sample Complaint Topic Trait Space visualizations, each corresponding to the first two principal components. Each gray dot corresponds to a complaint. Each red dot corresponds to a complaint about the concerned lender. The blue dot corresponds to a prospective borrower and is discussed in detail later in this paper.

3.7 Challenges in evaluation

The RMSE gives some indication of whether the predicted Complaint Topic Trait Space is an acceptable generalization of the data, but is otherwise of limited utility [8]. It is not a reliable metric in the case of predicting scores for a customer who interacts for another loan or with another lender in the future, or in the case of a new customer [33]. Nevertheless, our RMSE scores seem to suggest a generalization in the data compared with the maximum scores.

A major challenge in evaluating our work is that it is a case of unsupervised learning, and we do not have even a small labeled subset of the dataset. Given the large number of lenders, a controlled experiment would not be practical, so the best lender for a specific borrower — from a customer perspective — cannot be known with certainty or with a high confidence. While finding a true best lender is impractical, given that an alternative baseline is a random selection from a service quality perspective, we believe that recommendations to the customer (discussed in the next section) could be empowering even if the improvement (given that service quality is not the only factor) were incremental.

3.8 Making personalized recommendations

Table 4 shows samples visualizations of the topic trait embedding spaces. For each of the two techniques (Funk SVD and hybrid), and for each type of loan, the figure shows the spaces created from the complaint data of two (anonymized) banks or lenders. The blue dots correspond to a hypothetical prospective mortgage loan borrower (one of the authors!) using tweets from the hypothetical prospective borrower's Twitter handle. The Python-Twitter library [4] was used to fetch tweets (retweets are excluded) and infer the proxy personality traits. The hypothetical proxy borrower is asked to give relative importance scores for two randomly selected topics from the set of interpretable topics. Modified Collaborative Topic Regression is used to find the predicted 160 scores (predicted scores for the 10 proxy personality traits and two complaint topics are also included). The first two principal components are used to visualize the hypothetical proxy borrower as the blue dot.

We hypothesize that personalized recommendations can be made by finding a lender with the least number of complaints in the neighborhood of the prospective borrower, normalized by number of customers for that lender. The U.S. Federal Reserve publishes the number of branches of various lenders, and we use the number of branches as a proxy for the number of retail customers. One significant limitation of this proxy is that it ignores other (non-branch) channels of lending, and moreover assumes that all the considered lenders make loans of different types in a similar proportion, which is a significant approximation [52]. Unfortunately, the names of the lenders in the CFPB dataset and in the Federal Reserve Statistical Release [47] are often not identical, so we used the Levenshtein distance [43] to map between lender names. In the case where a lender is missing, we conservatively consider that lender to have only one branch.

These assumptions and approximations are significant limitations to the preliminary work we present here and which could be addressed with adequate time and effort. Because of this and because our intent is to present a methodology rather than a production ready system, we present anonymized lender names in Table 4.

4 RESULTS

We use the Freeman and Halton Exact Test to very conservatively evaluate our proposed methodology. Halkidi et al [26] say that for an external criteria test the null hypothesis is that the dataset is randomly structured. We apply external criteria to the Complaint Topic Trait Space at the lender level. It is feasible that the sub-space of interest has distinctive frequency distributions across lenders. Besides, groups of lenders may be similar at the aggregate level. We consider only the first principal component of the Complaint Topic Trait Space. We discretized the first principal component into twenty discrete levels. Then we find the frequency distribution based on the discrete level corresponding to each complaint. We did not consider other principal components, so our application of external validation is a very conservative test. In the case of the Freeman and Halton Exact Test, the alternative hypothesis is about general (and not linear) association. Table 5 presents p-values of the Freeman and Halton Exact Test for Complaint Topic Trait Spaces specified by the Funk SVD modified Collaborative Topic Regression and the Hybrid modified Collaborative Topic Regression in the case of the three types of loans.

According to the results, at an aggregate level and at a 0.20 level of significance, the Complaint Topic Trait Space is distinctive for the lenders in the case of payday loans. However, we cannot make such an inference in the case of mortgage loans and student loans. It is interesting that the p-values for the Complaint Topic Trait Space built using the Hybrid method are lower in the case of mortgage

Type of loan	User layers	Item layers	Regularization	Learning rate	Iterations	Batch size
Mortgage	(12773,64), (64,16), (16,4), (4,4)	(160,4), (4,4)	0.01	0.01	1	12773
Student	(2699,16), (16,4), (4,4)	(160,4), (4,4)	0.01	0.01	1	2699
Payday	(2699,16), (16,4), (4,4)	(160,4), (4,4)	0.01	0.01	1	2699

Table 2: Parameter values used for topic regression.

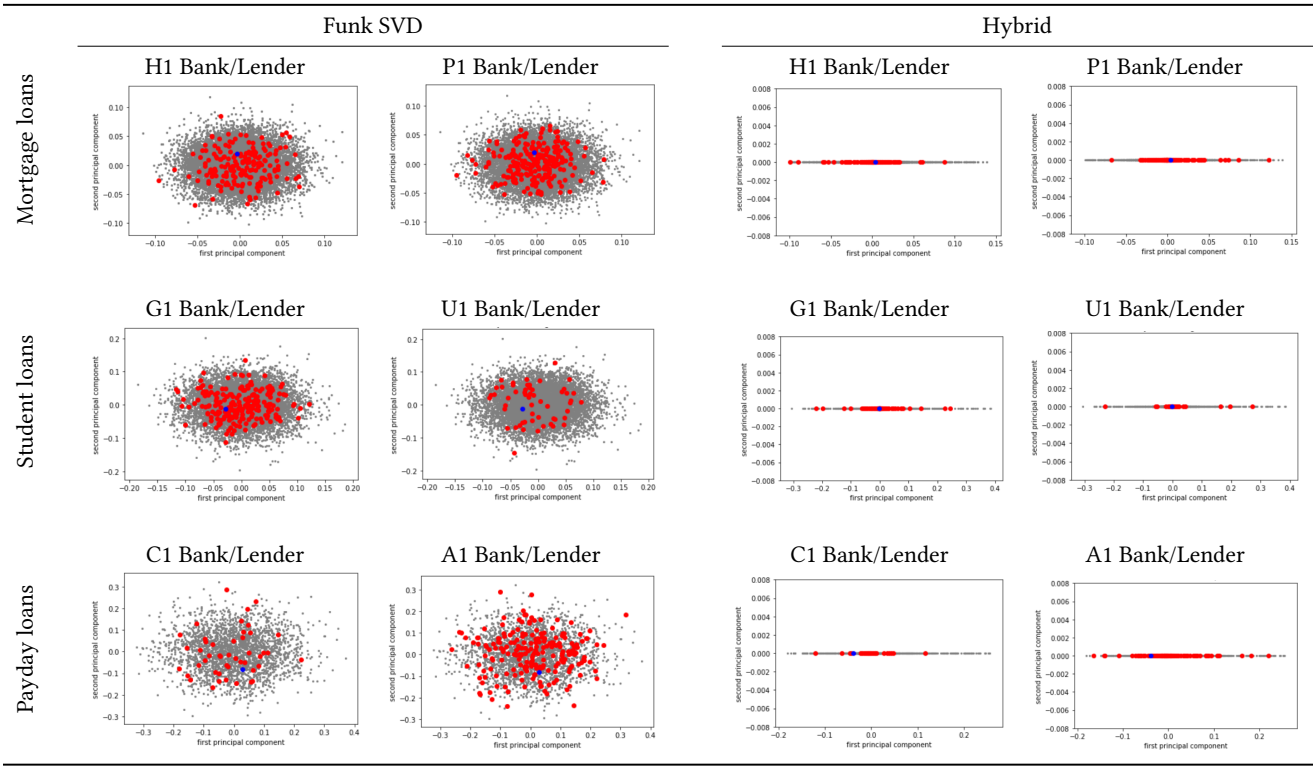


Table 4: Visualizations of complaint topic trait spaces. For each technique (Funk SVD and hybrid) and for each type of loan (mortgage, student loan, payday loan), we show a 2d projection of the embedding space for two sample (anonymized) lenders.

Loan type	Sample size	# samples	Space	p-value
Mortgage	12,772	1,000,000	Funk SVD	0.8029 ± 0.0010
			Hybrid	0.4962 ± 0.0013
Student	8,687	200,000	Funk SVD	0.6177 ± 0.0028
			Hybrid	0.5380 ± 0.0029
Payday	2,698	200,000	Funk SVD	0.0985 ± 0.0017
			Hybrid	0.1593 ± 0.0021

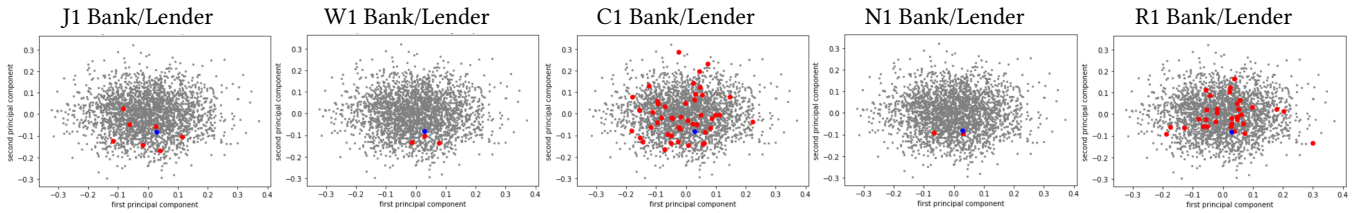
Table 5: Results of topic trait space analysis.

loans and student loans, which suggests that perhaps additional experimentation of the neural network topology and hyperparameters could lead to more distinctive spaces for the lenders (at an aggregate level). As discussed above, this test is very conservative.

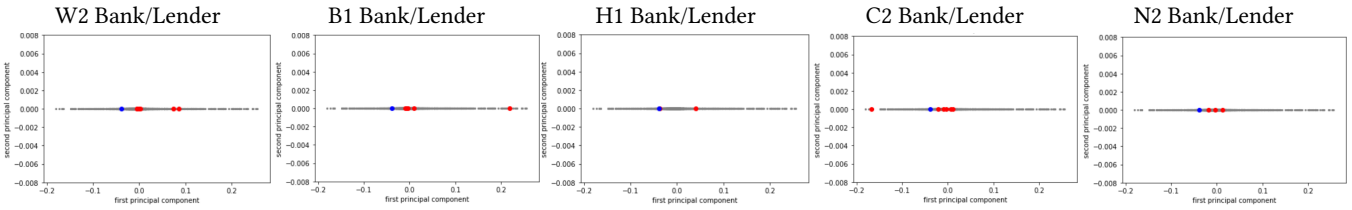
Sample recommendations using this space are shown in Table 6.

5 DISCUSSION

New datasets and open-source algorithms have made it possible to study complaints against financial institutions at a large scale. In particular, the availability of a cross-domain word embedding implementation like Vecmap as an open source project has been an essential tool that has made it — in conjunction with TREC 2011 Microblog Dataset and the WWBP — much easier to explore the utility of machine learning on a large complaints dataset from a customer's perspective. Interpretability of topics is a very significant problem and the use of modified Collaborative Topic Regression to meaningfully use a mix of interpretable and non-interpretable topic scores makes it possible to go beyond merely predicting based on personality traits alone. An alternative approach would have been to ignore an explicit consideration of complaint topics and instead use only personality trait embeddings [44]. This would require labeled data and would be an oversimplification, ignoring variations among people with similar personality traits but different appreciation of pain points. Domain adaptation in the case of such



(a) Funk SVD



(b) Hybrid

Table 6: Sample recommendations using the complaint topic space with Funk SVD (top) and the Hybrid method (bottom).

an alternative method would have to be implemented in a different way, perhaps as suggested in Rieman [48].

The hybrid method for building a Complaint Topic Trait Space seems promising given the substantially lower Freeman and Halton test p-values than with the Funk SVD for mortgage and student loans. While the use of complaints as a proxy for customer service is reasonable it has a limitation: it implicitly ignores positive actions by the lenders which may have a lesser but nevertheless important role in customer service. In case of a passive Twitter user, non-text inputs would be useful for inferring personality traits [53], however the inference of the personality proxy traits in such a case would be very different. Our method assumes that complaint users are representative of all retail customers.

Our work underscores that publicly available complaints datasets can be a gold mine for customers and service providers alike. Unfortunately, it has been reported that the CFPB complaints dataset may be withdrawn from the public [34]. In our view, in order to further empower the customer, more anonymized data (and not less) could be made available, including data about the originator of the loan, servicer of the loan, and each lender's number of retail customers for each type of loan by geography on a quarterly basis.

6 CONCLUSION AND FUTURE WORK

We proposed to analyze a publicly available complaints dataset from the retail customer's perspective. We showed that the Complaint Topic Trait Space is distinct for each lender in the case of payday loans while it was not distinct for mortgage and student loans. Some possible causes for this include a smaller intersection set of lenders originating loans and debt collectors, more frequent transfer of loans (among lenders), more heterogeneity among sub-products, and heterogeneity in service quality in different geographies. We proposed a method to make personalized recommendations about suitable lenders based on the customer service provided by those lenders. The visualizations of the Complaint Topic Trait Spaces suggest that the same method could be suitable in the case of mortgage loans and student loans depending on the location of the Twitter user in the Complaint Topic Trait Space. To confirm the same frequency distributions of the complaints in the Complaint Topic Trait Sub-spaces in the neighborhood of Twitter users and the distributions of the Complaint Topic Trait Sub-spaces for groups of lenders in the neighborhood of Twitter users, we will have to test using external criteria like the Freeman and Halton Exact tests.

In line with *primum non nocere*, given that the Freeman and Halton Exact test only tells us whether the structure of the dataset is not random (for low p-values), it would be prudent to use our proposed method only on the subset of payday loan lenders (and to mortgage loan lenders and student loans lenders based on future work) which in a specific customer's eyes are similar. In other words,at least initially, the Twitter user also ought to be asked for her shortlist of potential lenders from whom she is equally inclined to borrow and our proposed methodology ought to recommend from among that shortlist. In the future, lenders may analyze complaints against them across time to evaluate variation in quality of service. This may also be particularly useful for analyzing complaints before and after the launch of new products, and refinement of products based on such analyses. Use of anomaly detection (e.g. [40]) for finding both nominal and abnormal points in the Complaint Topic Trait Space could potentially give interesting insights. Metrics based on the proportion of nominal and abnormal points could also be useful.

Competitive pressures that could result from a customer focus on quality are desirable, and could help lenders move away from "a race to the bottom" based primarily on pricing (which also contributes— along with other factors— to unviable low interest rates which may be accompanied by asset price bubbles and hence a less stable economy). Future work could include use of the proposed method in the case of other publicly available complaints datasets, better interpretability of topics by multiple experts, use of bigrams and trigrams, experimentation with the hyperparameters and additional topologies for the hybrid method for building the Complaint Topic Trait Spaces, benchmarking of proxy personality trait scores, and

use of Facebook text [29] or an ensemble of tweets and Facebook posts. Mortgage loan sub-products could be analyzed separately. Federal student loans and private student loans could also be analyzed separately. Analysis may also be undertaken by geography. Under the assumption that it is the meanings of words that is relevant for association with a personality trait, an extension to other languages will also be feasible using Cross-Lingual word embeddings. Finally, our techniques could be extended and applied to other parts of the service economy.

7 ACKNOWLEDGMENTS

We acknowledge first hearing about the belief regarding the asymmetrical effect of loss/gain associated with pain/joy from A. V. Rajwade (1936-2018). We acknowledge the help of Maciej Kula about the appropriate class in the Spotlight library that could be modified for adding more layers to the neural network. This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute, and in part by the Indiana METACyt Initiative. The Indiana METACyt Initiative at IU was also supported in part by Lilly Endowment, Inc.

REFERENCES

[1] [n. d.]. https://github.com/artetxem/vecmap.
[2] [n. d.]. https://www.consumerfinance.gov/data-research/consumer-complaints/.
[3] [n. d.]. Port of Google's Language Detection library to Python. https://pypi.org/project/langdetect/.
[4] [n. d.]. python-twitter: A Python wrapper around the Twitter API. https://python-twitter.readthedocs.io/en/latest/.
[5] [n. d.]. TREC 2011 Twitter Collection Downloader. https://github.com/cmlonder/trec-collection-downloader.
[6] 2017. How To Choose the Right Bank for You. *ABC News* (2017).
[7] 2017. Spotlight. https://github.com/maciejkula/spotlight.
[8] 2018. Implicit and explicit feedback recommenders: And the curse of RMSE. https://resources.bibblio.org/hubfs/share/2018-01-24-RecSysLDN-Ravelin.pdf.
[9] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Annual Meeting of the Association for Computational Linguistics*.
[10] Ian Ayres, Jeff Lingwall, and Sonia Steinway. 2013. Skeletons In The Database: An Early Analysis Of The CFPB's Consumer Complaints. *Fordham Journal of Corporate & Financial Law* XIX (2013).
[11] Financial Stability Board. 2017. Artificial intelligence and machine learning in financial services: Market developments and financial stability implications. , 11-15 pages.
[12] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv:1607.04606* (2016).
[13] Consumer Financial Protection Bureau. 2015. Consumers' mortgage shopping experience: A first look at results from the National Survey of Mortgage Borrowers.
[14] Consumer Financial Protection Bureau. 2017. CFPB Summary of product and sub-product changes.
[15] Consumer Financial Protection Bureau. 2018. Buying a house: Tools and resources for homebuyers. https://www.consumerfinance.gov/owning-a-home/process/compare.
[16] Bureau Of Consumer Financial Protection. 2018. Complaint snapshot: Debt collection.
[17] Antal Buza and Piroska B. Kis. 2014. Instability Of Matrix Factorization Used In Recommender Systems. *Annales Univ. Sci. Budapest., Sect. Comp.* 42 (2014).
[18] Fabio Celli, Fabio Pianesi, David Stillwell, and Michal Kosinski. 2013. Workshop on Computational Personality Recognition: Shared Task. In *ICWSM Workshops*.
[19] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Neural Information Processing Systems*.
[20] Stacy Cowley. 2018. Consumer Bureau Looks to End Public View of Complaints Database. *New York Times* (2018).
[21] Deloitte. 2015. The power of complaints: Unlocking the value of customer dissatisfaction.
[22] Fabrizio Esposito, Anna Corazza, and Francessco Cutugno. 2016. Topic Modeling with Word Embeddings. In *Third Italian Conference on Computational Linguistics*.
[23] Pamela Foohey. 2017. Calling On The CPFB For Help: Telling Stories And Consumer Protection. In *80 Law & Contemporary Problems (Forthcoming) Indiana Legal Studies Research Paper No. 356*.
[24] Andreas Fuster, Matthew Plosser, and James Vickery. 2018. Does CFPB Oversight Crimp Credit? . *Federal Reserve Bank of New York Staff Reports* 857 (2018).
[25] Boris A. Galitsky and Josep Lluis de la Rosa. 2011. Learning Adversarial Reasoning Patterns in Customer Complaints. In *AAAI Workshop on Applied Adversarial Reasoning and Risk Modeling*.
[26] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. 2001. On Clustering Validation Techniques. *Journal of Intelligent Information Systems* 17, 2/3 (2001).
[27] Matthew Honnibal and Mark Johnson. 2015. An Improved Non-monotonic Transition System for Dependency Parsing. In *EMNLP*. 1373–1378.
[28] Nicholas Hug. 2017. Surprise, a Python library for recommender systems. http://surpriselib.com.
[29] Kokil Jaidka, Sharath Chandra Guntuku, Anneke Buffone, H. Schwartz, and Lyle Ungar. 2018. Facebook vs. Twitter: Differences in Self-disclosure and Trait Prediction. In *ICWSM*.
[30] J.D. Power Ratings. 2016. Buyer's Remorse Is Relatively High despite Rising Satisfaction.
[31] J.D. Power Ratings. 2017. Secret to Nationwide, Multiproduct Success in Retail Banking: Not Making Mistakes.
[32] D. Kahneman and A. Tversky. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47, 4 (1979), 263âĂŞ291.
[33] Frank Kane. [n. d.]. Building Recommender Systems with Machine Learning and AI. https://www.udemy.com/building-recommender-systems-with-machine-learning-and-ai/.
[34] Ted Knutson. 2018. CFPB Chief: I could make consumer complaints secret. *Forbes* (2018).
[35] KPMG. 2017. KPMG Client Alert, America's FS Regulatory Center of Excellence: Complaints Monitoring and Risk Management.
[36] Maciej Kula. 2016. Hybrid Recommender Systems at PyData Amsterdam 2016. https://speakerdeck.com/maciejkula/hybrid-recommender-systems-at-pydata-amsterdam-2016?slide=21.
[37] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised Machine Translation Using Monolingual Corpora Only. In *ICLR*.
[38] Quanzhi Li, Sameena Shah, Xiaomo Liu, and Armineh Nourbakhsh. 2017. Word Embeddings Learned from Tweets and General Data. In *ICWSM*.
[39] Jimmy Lin. [n. d.]. Twitter tools. https://github.com/lintool/twitter-tools.
[40] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2012. Isolation-based Anomaly Detection. *ACM Transactions on Knowledge Discovery from Data* 6, 1 (March 2012).
[41] Richard McCreadie, Ian Soboroff, Jimmy Lin, Craig Macdonald, Iadh Ounis, and Dean McCullough. 2012. On building a reliable Twitter corpus. In *SIGIR*.
[42] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv* (2013).
[43] David Necas. [n. d.]. python-Levenshtein:Python extension for computing string edit distances and similarities. https://pypi.org/project/python-Levenshtein/.
[44] Yair Neuman and Yochai Cohen. 2014. A Vectorial Semantics Approach to Personality Assessment. *Scientific Reports* 4 (2014).
[45] Rehurek Radim and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Workshop on New Challenges for NLP Frameworks*.
[46] Lucia Rahilly. 2018. *McKinsey on Risk*.
[47] Federal Reserve Statistical Release. 2018. Large Commercial Banks: Insured U.S.-Chartered Commercial Banks that have Consolidated Assets of $300 million or more, ranked by Consolidated Assets.
[48] Daniel Rieman, Kokil Jaidka, H. Schwartz, and Lyle Ungar. 2017. Domain Adaptation from User-level Facebook Models to County-level Twitter Predictions. In *International Joint Conference on Natural Language Processing*.
[49] Tom Sabo. 2017. *Applying Text Analytics and Machine Learning to Assess Consumer Financial Complaints*. Technical Report. SAS Institute.
[50] H. Schwartz, Johannes Eichstaedt, Margaret Kern, Lukasz Dziurzynski, Stephanie Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin Seligman, and Lyle H. Ungar. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLOS ONE* 8, 9 (September 2013).
[51] Laura Shin. 2014. How To Choose A Bank Account: 10 Things To Look For. *Forbes* (2014).
[52] Trefis Group. 2017. A breakdown of the loan portfolios of the largest U.S. banks.
[53] Svitlana Volkova, Yoram Bachrach, and Benjamin Van Durme. 2016. Mining User Interests to Predict Perceived Psycho-Demographic Traits on Twitter. In *Second International Conference on Big Data Computing Service and Applications*.
[54] Chong Wang and David Blei. 2011. Collaborative Topic Modeling for Recommending Scientific Articles. In *KDD*.
[55] Yandex. [n. d.]. Big Data Applications: Machine Learning at Scale, Coursera. https://www.coursera.org/lecture/machine-learning-applications-big-data/recsys-mf-ii-Bq6m2.