



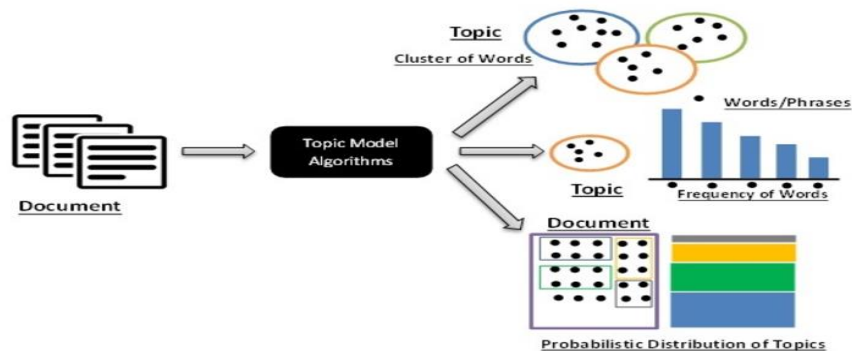
Complaint Text Categorization using different topic modelling approaches (LDA, BERT)

October 26, 2020

Topic Modeling on customer complaints text

Business Case

- Customer centric business, it is very much required to ensure they are acting in their customers' best interest, which includes the way in which they address their customer complaints.
- Text mining and topic modeling techniques are involved to examine unstructured text present within the complaints texts and to classify them under most frequent and meaningful topics through unsupervised methods.
- Topic modeling in complaints text can help business to understand what the frequently occurring customer dissatisfaction areas are and where improvement in customer servicing will increase business outcome and customer satisfaction.
- At call centres, agent manually assign category/topic as per customer issue. There remain chances of misassignment of Complain/Issue category.
- Here our work is to automate this category assignment.



Solution

- Topic modeling is an unsupervised machine learning approach that identifies **'themes' within text data**
- Here we have build a topic model on sample complaint text of "Mortgage" product (Data source: CFPB). Then we run different unsupervised topic modelling approaches (kMeans, LDA on both TFIDF & BERT vectors), at first to define total optimal no of complaint category and then assign the most suitable issue category to the new complain text in "Morgage" product.
- LDA(Latent Dirichlet Allocation) is a topic modeling algorithm that has been included in gensim. It classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions.
- BERT sentence embedding model embed text into a vector space where the vectors capture the contextual meaning of sentences.

Benefits

- ✓ Customer negative feedback/complaint data is categorized and convened into countable topics and servicing areas that can be easily comprehended by business and complaint management team.
- ✓ Gain valuable insights on the historic trend of issues occurring in customer servicing/support activities.
- ✓ Categorization and segmentation of topic are detected through unsupervised and help business to take more active decision for a particular segment.
- ✓ Trends can be derived, which segment of topic is trending over a time.
- ✓ Identify fragments, where more keen decision can be taken.
- ✓ More effective management decision can be taken and weak areas can be identify.

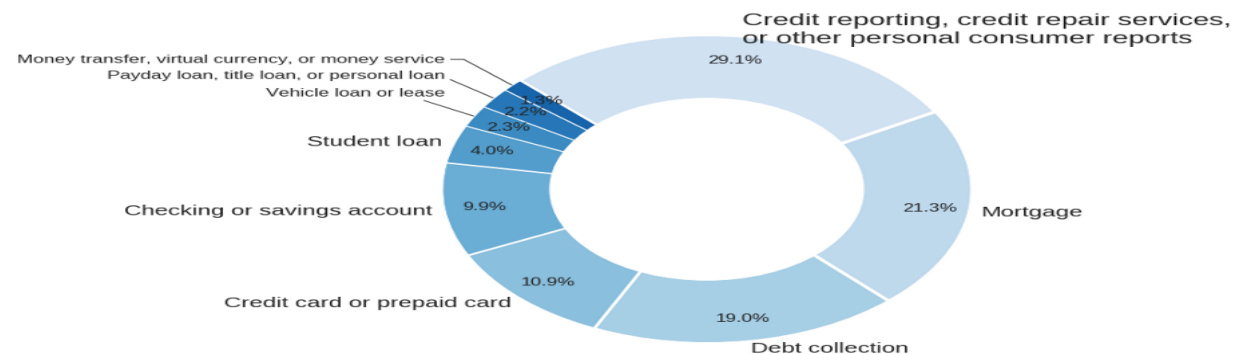
CFPB(Consumer Financial Protection Bureau) database

CFPB publishes data on consumer complaints about financial products and services. This data is captured either after the company responds to the complaint, or after 15 days.

What are customers complaining about?

Complaints handled by the Consumer Financial Protection Bureau

Between December 1, 2011 and June 22, 2019, the CFPB received approximately 1,315,085 customer complaints. Here's how they break down by category.



Product	count	mean	std	min	25%	50%	75%	max
Debt collection	47,915	815	829	5	301	557	1,027	25,491
Mortgage	36,582	1,470	1,154	13	622	1,129	2,033	31,369
Credit reporting	31,592	750	721	11	275	509	960	4,171
Credit card	18,842	1,126	895	14	481	861	1,503	4,029
Bank account or service	14,888	1,243	954	9	533	954	1,683	5,151
Credit reporting, credit repair services, or other personal consumer reports	14,671	868	947	12	338	621	1,083	21,112
Student loan	13,304	1,183	1,001	6	497	890	1,554	17,604
Consumer Loan	9,474	1,108	911	18	437	821	1,496	4,045

Mortgage Complaint Text descriptive statistics

Mortgage



Related issues

Loan servicing, payments, escrow account



Loan modification, collection, foreclosure



Top 3 Organizations

out of 2355 comments



276 (11.72%)



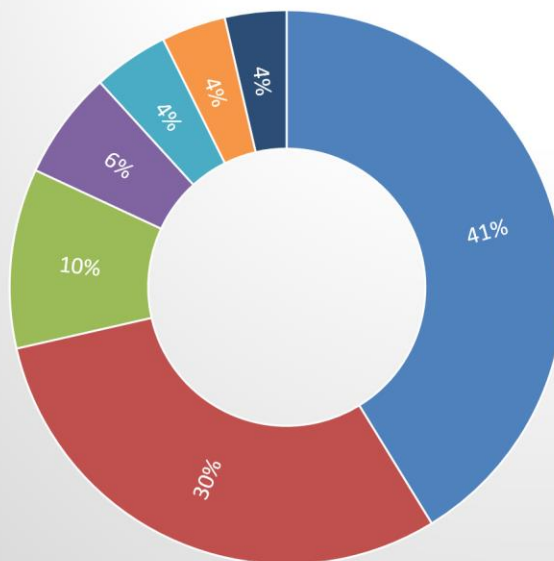
240 (10.19%)



ditech®

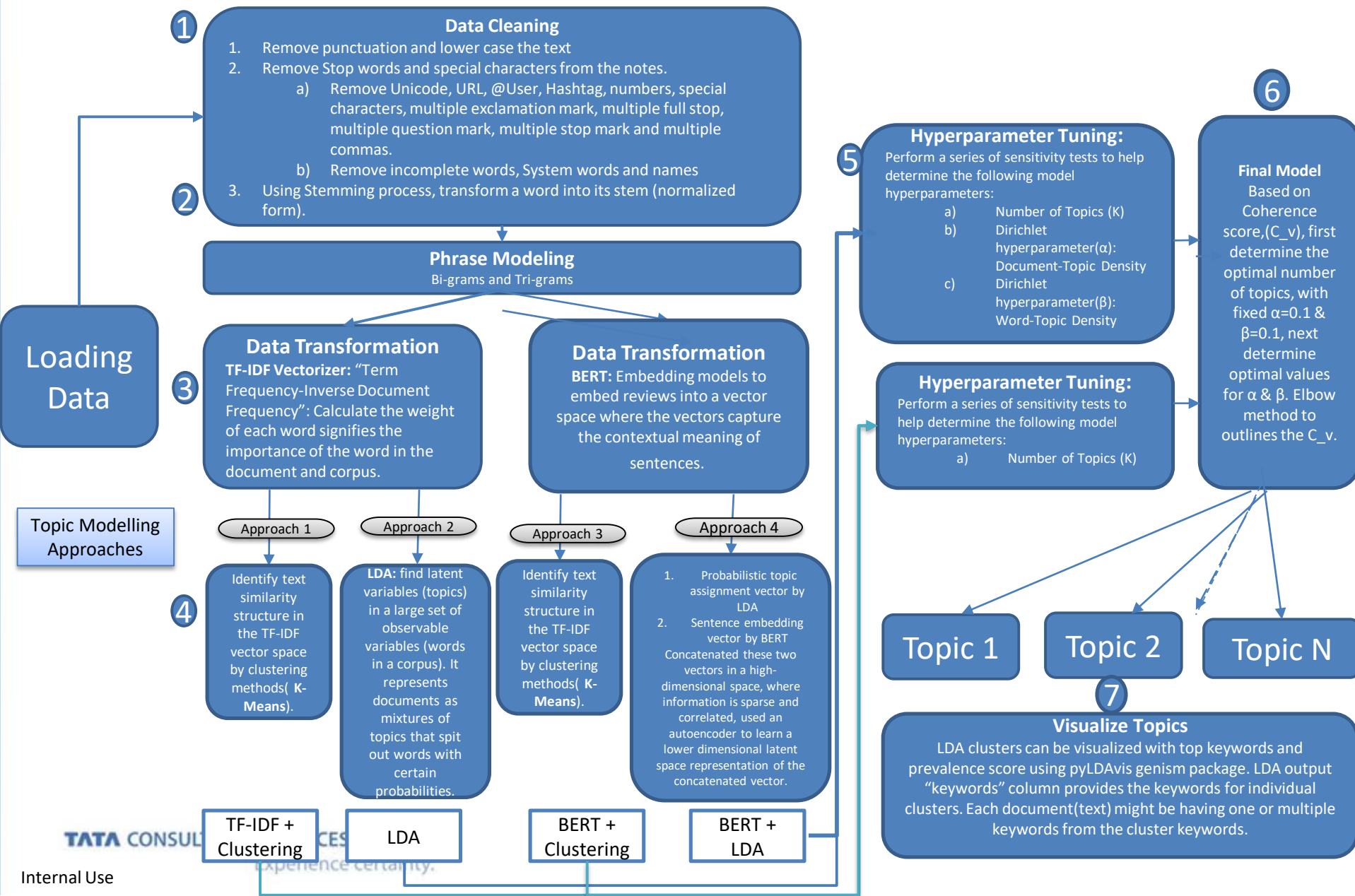
170 (7.22%)

Mortgage complain text 'Issue' category dist%



- Loan servicing, payments, escrow account
- Loan modification, collection, foreclosure
- Application, originator, mortgage broker
- Settlement process and costs
- Trouble during payment process
- Struggling to pay mortgage
- Credit decision / Underwriting

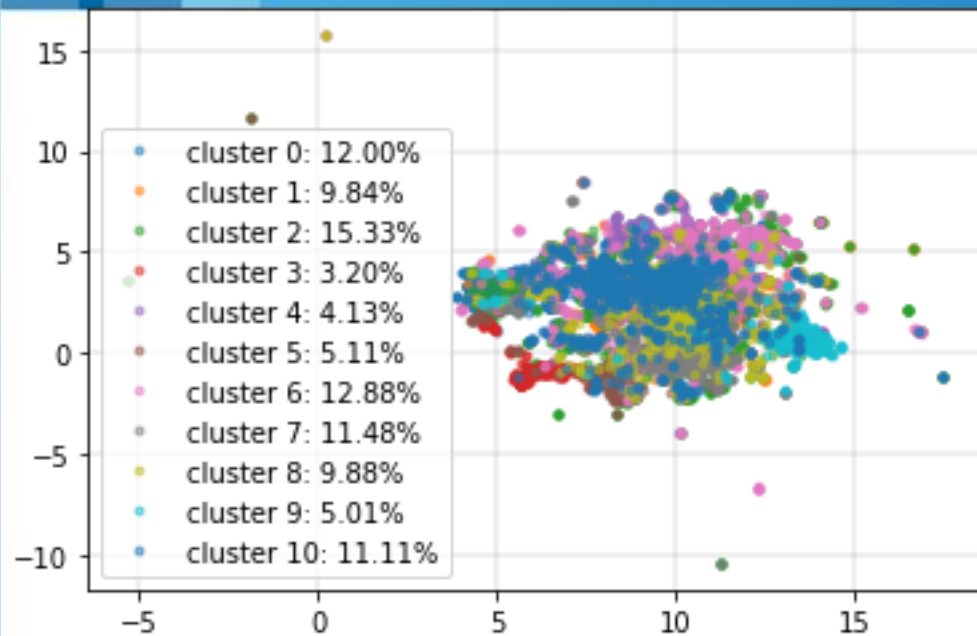
Our Approaches



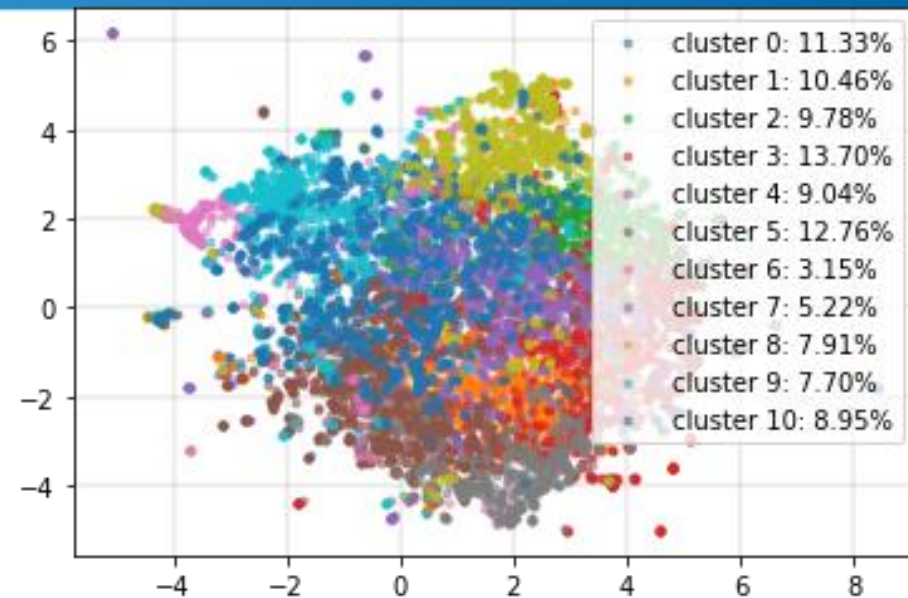
Result Comparison

Evaluation	Approaches			
Metric/Method	TF-IDF + Clustering	LDA	BERT + Clustering	BERT + LDA
Coherence Value	0.266975732	0.301286374	0.395795941	0.436755
Silhouette score	0.008472792	0.101698765	0.030916205	0.15714453

Clustering Results

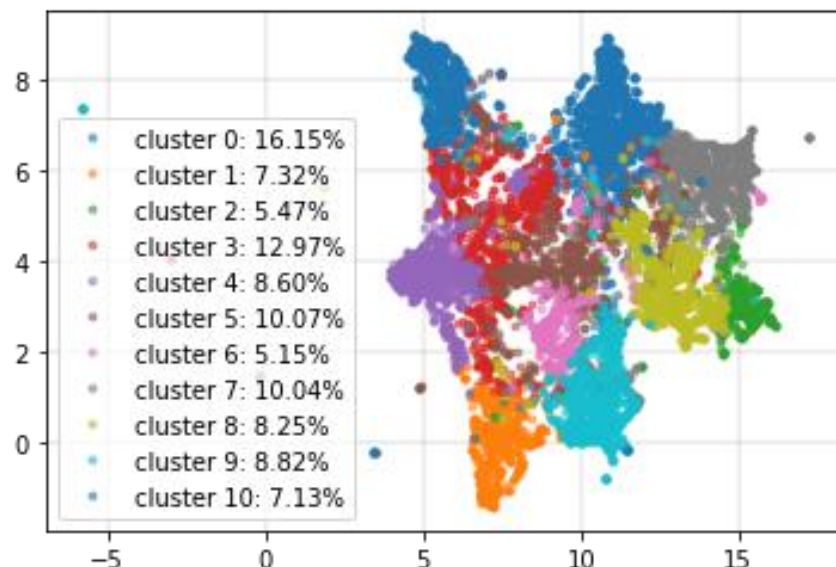


TF-IDF RESULTS CLUSTERS



LDA RESULTS CLUSTERS

BERT+LDA RESULTS
CLUSTERS

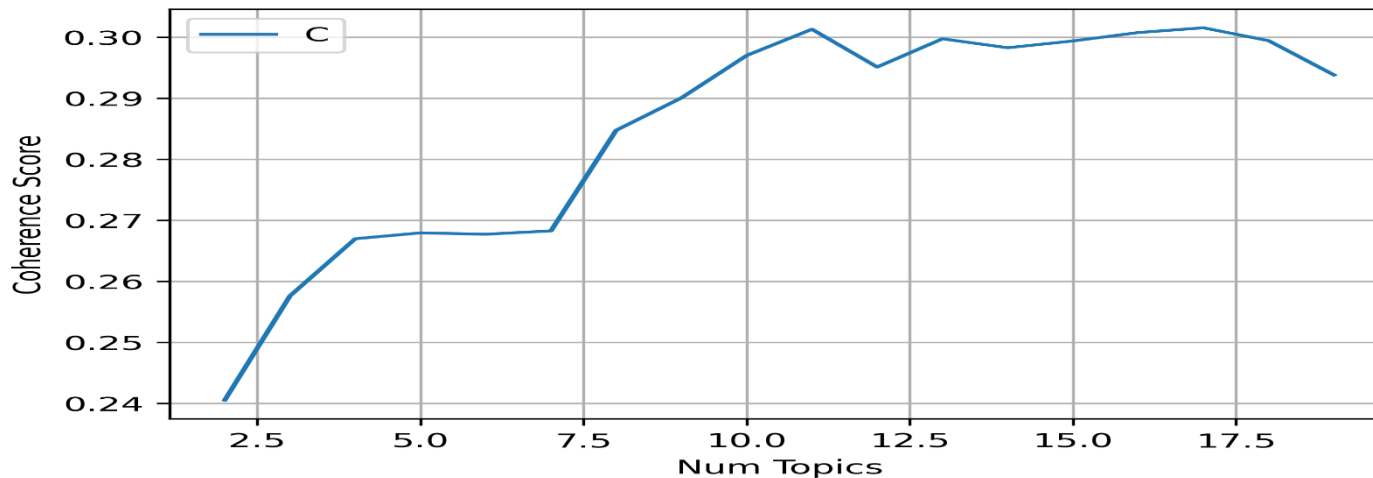


By combining LDA, BERT, and clustering, we can keep the semantic information and create Contextual Topic Identification. Below is the result, where clusters are balanced and quite separated.



LDA+BERT (contd..)

LDA hyperparameter tuning output for various n_topics.



For $\alpha=0.3$ $\beta=0.9$ & $n_topic=11$ we are getting high Coherence score

LDA output topics distribution vs actual source issue category

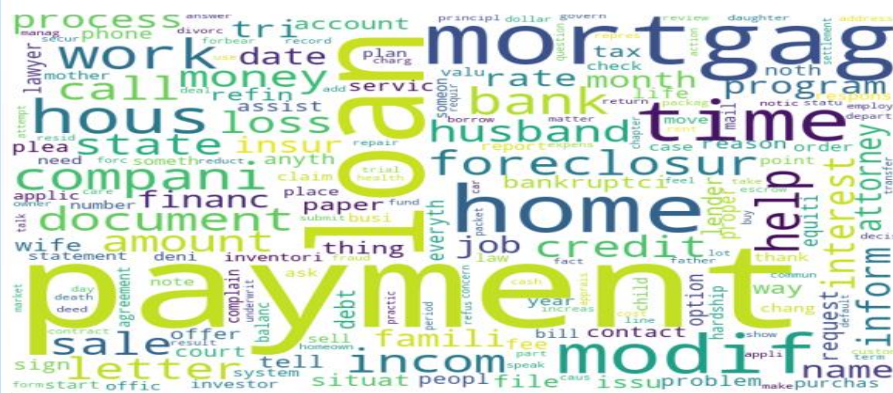
Mortgage Complaint Text Issue Type	LDA_Topic_1	LDA_Topic_2	LDA_Topic_11	LDA_Topic_3	LDA_Topic_4	LDA_Topic_5	LDA_Topic_6	LDA_Topic_7	LDA_Topic_8	LDA_Topic_9	LDA_Topic_10	Grand Total
Application, originator, mortgage broker		495	223									718
Collection				681				433				1,114
Credit decision / Underwriting		571				287	207					1,065
Escrow account			812			751		476	474			2,513
Foreclosure	520						521					1,041
Loan modification					321				318	416	186	1,241
Loan servicing	808			931						471	275	2,485
Payments					350		597					947
Settlement process and costs		198	187									385
Struggling to pay mortgage						48			39			87
Trouble during payment process	73					52		25	34			184
Grand Total	1,401	1,264	1,222	1,612	671	1,138	1,325	934	865	887	461	11,780

Internal Use												
LDA Cluster Output												
Mortgage Complaint Text Issue Type	LDA_Topic_1	LDA_Topic_2	LDA_Topic_11	LDA_Topic_3	LDA_Topic_4	LDA_Topic_5	LDA_Topic_6	LDA_Topic_7	LDA_Topic_8	LDA_Topic_9	LDA_Topic_10	Grand Total
Application, originator, mortgage broker		495	223									718
Collection				681				433				1,114
Credit decision / Underwriting		571				287	207					1,065
Escrow account			812			751		476	474			2,513
Foreclosure	520						521					1,041
Loan modification					321				318	416	186	1,241
Loan servicing	808			931						471	275	2,485
Payments					350		597					947
Settlement process and costs		198	187									385
Struggling to pay mortgage						48			39			87
Trouble during payment process	73					52		25	34			184
Grand Total	1,401	1,264	1,222	1,612	671	1,138	1,325	934	865	887	461	11,780

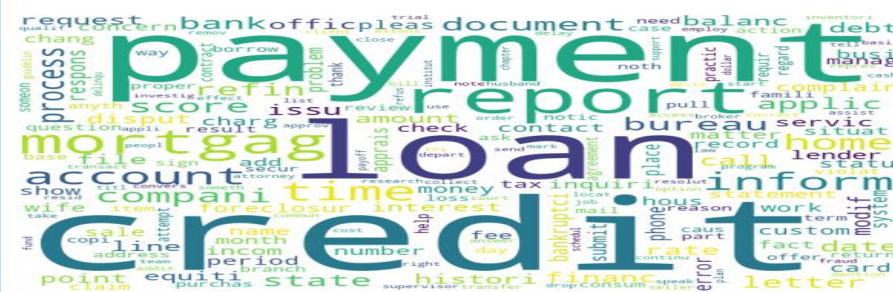
[illegible]

Most frequent keywords are 'Account' 'Mortgage' which derived it as **'Application Originator Mortgage broker'** & 'Credit' 'Report' derives it as **'Credit decision / Underwriting'** issue category.

[illegible]



Most frequent keywords are 'Loan' 'Help' 'Service' 'Call' 'Statement' which derived it as **'Loan Servicing'** & 'Foreclosure' 'Refinance' 'Inform' derives it as **'Foreclosure'** issue category



Most frequent keywords are 'Account' 'Mortgage' which derived it as **'Application Originator Mortgage broker'** & 'Credit' 'Report' derives it as **'Credit decision / Underwriting'** issue category



Most frequent keywords are 'Payment' 'Time' 'Statement', 'Money', 'Balance', 'Amount', 'Bill' which derived it as **'Payment'** & 'Loan' 'Modification' 'Letter' , 'Bill', 'Foreclosure', 'Document', 'Charge', 'Refinance', 'Fee' derives it as **'Loan Modification'** issue category

To Do(Scope for improvement)

- 1) For such type of big document text(containing mixture keywords from multiple topics), it is hard to assign any particular target topic by any unsupervised technique. To improve topic modelling process, in input we can apply advanced 'Text Summarize' approach to reduce the input text length and mixed context words.
- 2) Using advanced 'Sentiment Analysis' technique we can filter out non-complain text in input.



Appendix

Data Preprocessing(Stop-words)

- **Stop-words are those words from non-linguistic view which do not carry information**
 - ...they have mainly functional role
 - ...we have removed them to help the methods to perform better
- **Natural language dependent** – examples:
 - English: A, ABOUT, ABOVE, ACROSS, AFTER, AGAIN, AGAINST, ALL, ALMOST, ALONE, ALONG, ALREADY, ALSO, ...

Stemming & Lemmatization

- Different forms of the same word are usually problematic for text data analysis, because they have different spelling and similar meaning (*e.g. learns, learned, learning,...*)
- **Stemming** is a process of transforming a word into its stem (*normalized form*)
- **Lemmatization:** The lemma (root word) has been identified from the individual tokens using NLTK wordnet lemmatizer based on the significance of the word (NLTK pos-tag) in the corresponding sentence.

Phrases in the form of frequent N-Grams

- Simple way for generating phrases are frequent n-grams:
 - **N-Gram** is a sequence of n consecutive words (e.g. "machine learning" is 2-gram)
 - "**Frequent n-grams**" are the ones which appear in all observed documents MinFreq or more times.

LDA: A very simple example

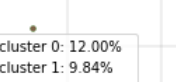
Consider these 5 “documents”:

- I like to eat broccoli and bananas.
- I ate a banana and spinach smoothie for breakfast.
- Chinchillas and kittens are cute.
- My sister adopted a kitten yesterday.
- Look at this cute hamster munching on a piece of broccoli.

If asked for 2 topics, LDA might produce:

- **Sentences 1 and 2:** 100% Topic A
- **Sentences 3 and 4:** 100% Topic B
- **Sentence 5:** 60% Topic A, 40% Topic B
- **Topic A:** 30% broccoli, 15% bananas, 10% breakfast, 10% munching, ... (at which point, you could interpret topic A to be about food)
- **Topic B:** 20% chinchillas, 20% kittens, 20% cute, 15% hamster, ... (at which point, you could interpret topic B to be about cute animals)

Note: in order to do this, LDA assumes prior probability distribution and then iteratively generates topics using a bag-of-words representation

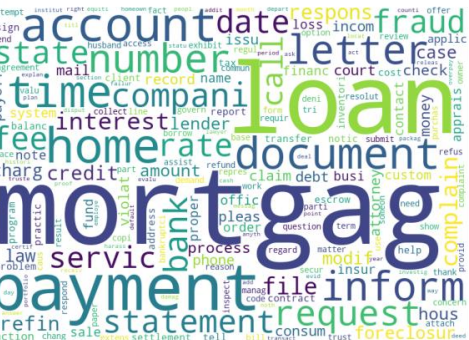


A scatter plot showing 10 clusters of data points. The x-axis ranges from -5 to 15, and the y-axis ranges from 0 to 10. The clusters are represented by different colors and are mostly concentrated between x=5 and x=15. A legend on the left lists the clusters and their percentages:

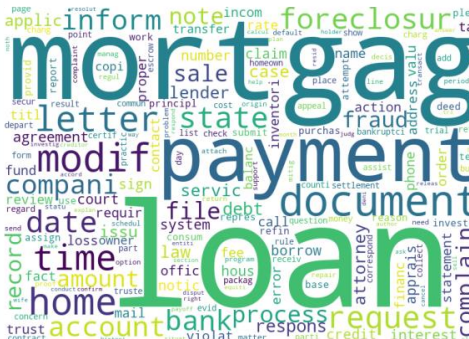
- cluster 0: 12.00%
- cluster 1: 9.84%
- cluster 2: 15.33%
- cluster 3: 3.20%
- cluster 4: 4.13%
- cluster 5: 5.11%
- cluster 6: 12.88%
- cluster 7: 11.48%
- cluster 8: 9.88%
- cluster 9: 5.01%
- cluster 10: 11.11%

[illegible][illegible]

Cluster 11

[illegible][illegible]

Cluster 6

[illegible][illegible]

payment collect address period
house inform
system history lender equities attempt
status court bus correct train
sign credit debt point call
refin move husband
home
account
bankrupt
interest bank
rate inquiry
claim
dispute
money make
copi show
action
file wife
balanc
process
loan
dropt
compani
servic
consum
manip
complain
review
request
burea
chang
scor
add not
effect
record
thing
foreclosur
mortgag
disapp
pleas
mount
month
term
plain
in

Cluster 10

LDA Visualization(pyLDAvis package)

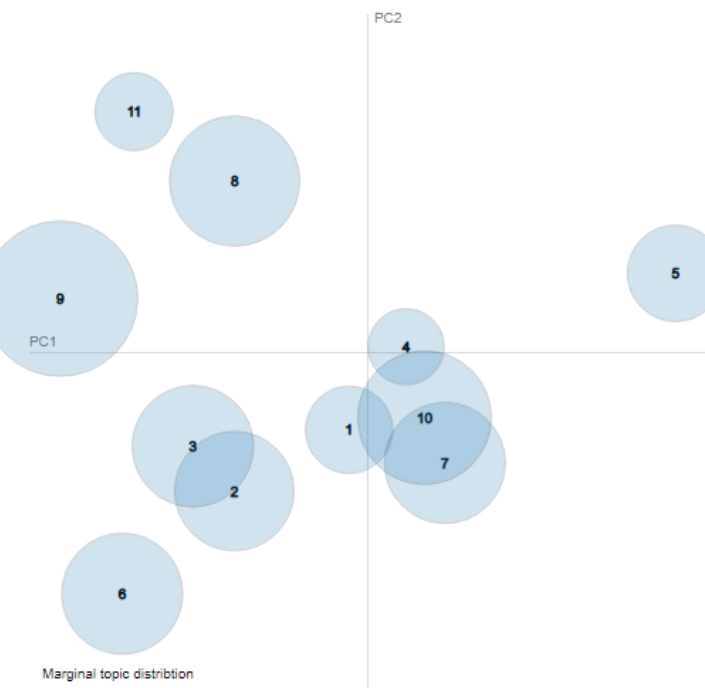
Selected Topic: 0

Slide to adjust relevance metric:⁽²⁾

$\lambda = 1$

0.0 0.2 0.4 0.6 0.8 1.0

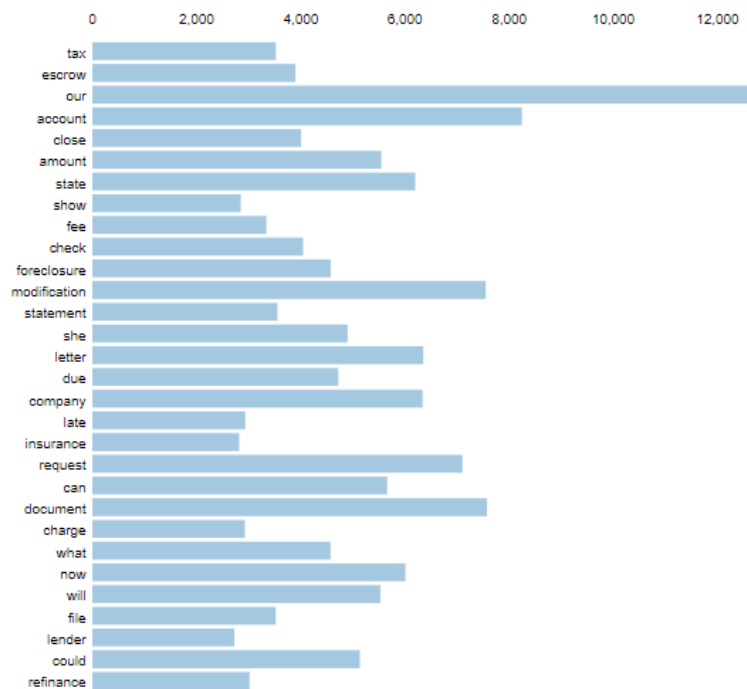
Intertopic Distance Map (via multidimensional scaling)



Marginal topic distribution



Top-30 Most Salient Terms⁽¹⁾



Overall term frequency

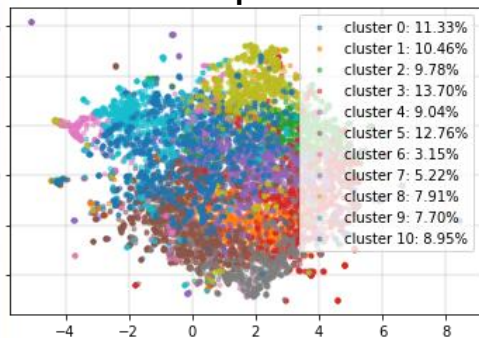
Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))]; see Chuang et. al (2012)

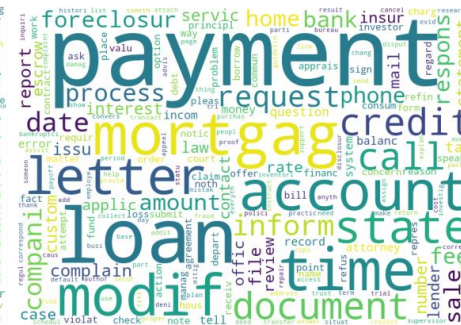
2. relevance(term w | topic t) = $\lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$; see Sievert & Shirley (2014)

BERT + Clustering

Using sentence embedding models (BERT) to embed text into a vector space where the vectors capture the contextual meaning of sentences.



Cluster 1



Cluster 2



Cluster 3



Cluster 4



Cluster 5



Cluster 6



Cluster 7



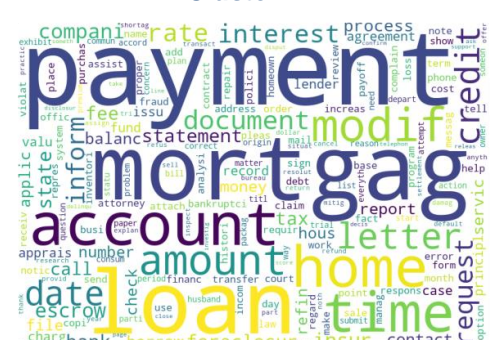
Cluster 8



Cluster 9



Cluster 10



Cluster 11

