UNIVERSITEIT GENT

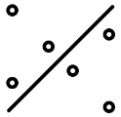
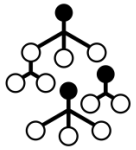
Machine Learning

Final presentation

GROUP 9

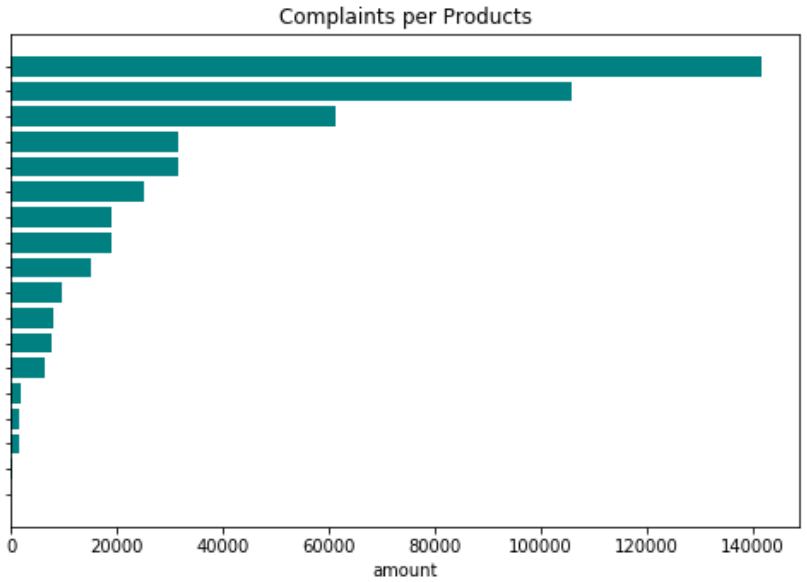
BRYAN VAN HUYNEGHEM – ANOEK STRUMANE

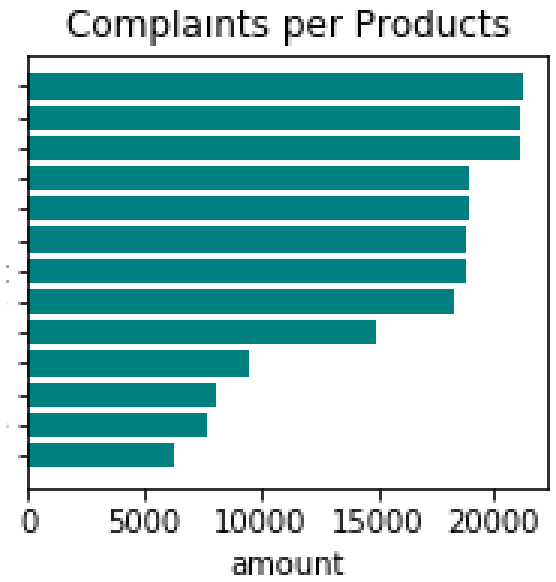
Table of contents

- ***Sprint 1***
  - Exploratory Data Analysis (EDA)
  - Data Pre-processing & Cleaning
  - Supervised Learning Algorithms: **SVM**  and **RF** 
- ***Sprint 2***
  - Finding a different classification
  - Exploring new word embedding methods (**Doc2Vec**)
  - Unsupervised Learning Algorithms: **K-Means**, **DBSCAN**, **K-Medoids** 
- ***Sprint 3***
  - Additional pre-processing
  - Unsupervised Learning Algorithm (Topic Modelling): **LDA** 
  - Supervised Learning Algorithm (Deep Neural Network): **BERT** 

EDA

- Imbalanced dataset





… still not perfect, but slightly better.
More improvements throughout sprints.

- Other issues:
  - Scarce data for some products
  - Unique columns / General columns
  - Narratives: need to be cleaned…

  - Removed complaints about certain products (== balancing)
  - Removed unique columns / General columns
  - Cleaned narratives

Data Pre-processing & Cleaning

'On XX/XX/XXXX I signed a car loan agreement to finance my 2008 XXXX XXXX XXXX car. \nThe loan amount was {$5400.00}, that is the amount that was financed. The interest was 8.54 %. The total amount to be paid back was {$6500.00}, in 48 months payments. The amount per month was {$130.00}. The payments began on XX/XX/XXXX and were to end on XX/XX/XXXX. Fortunately I made some pay ments that were much more than the monthly required amount of {$130.00}. That means that the final payment is now to be well be fore XX/XX/XXXX. In fact I did my last payment in XX/XX/XXXX. In addition I have mailed them a final payment of {$180.00} ({$5 0.00} deferral fee and {$130.00} for the payment I skipped XX/XX/XXXX). \nDuring the term of loan I received no communication, no monthly account statement and no notice of late payments. Also I have been making my payment on time every month with the ex

Cleaning

'on i signed a car loan agreement to finance my car the loan amount was that is the amount that was financed the interest was t he total amount to be paid back was in months payments the amount per month was the payments began on and were to end on fortun ately i made some payments that were much more than the monthly required amount of that means that the final payment is now to be well before in fact i did my last payment in in addition i have mailed them a final payment of deferral fee and for the paym ent i skipped during the term of loan i received no communication no monthly account statement and no notice of late payments a lso i have been making my payment on time every month with the exception of the skipped payment in i have called the bank to le

Lemmatization

- Removing plurals, conjugations → smaller BoW

Complaint

CountVectorizer

TF-IDF

Dimensionality Reduction

$$\begin{bmatrix} & T_1 & T_2 & \cdots & T_t \\ D_1 & w_{11} & w_{21} & \cdots & w_{t1} \\ D_2 & w_{12} & w_{22} & \cdots & w_{t2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ D_n & w_{1n} & w_{2n} & \cdots & w_{tn} \end{bmatrix}$$

Document-Term Matrix

Results with SVM: ***narrative*** only

- Unbalanced dataset (as an experiment)

Run	PC	kernel	accuracy
1	5	linear	0.47
2	50	linear	0.63
3	5	rbf	0.51

- Balanced dataset

Run	PC	kernel	accuracy
4	15	linear	0.47
5	30	linear	0.52
6	50	linear	0.55
7	100	linear	0.58

Results with SVM: ***narrative*** only

	precision	recall	f1-score	support
Bank account or service	0.17	0.58	0.26	1081
Checking or savings account	0.77	0.51	0.61	7168
Consumer Loan	0.13	0.53	0.20	562
Credit card	0.46	0.51	0.49	4337
Credit card or prepaid card	0.45	0.53	0.49	3957
Credit reporting	0.54	0.52	0.53	4933
Credit reporting, credit repair services, or other personal consumer reports	0.46	0.42	0.44	5635
Debt collection	0.78	0.55	0.64	7609
Money transfer, virtual currency, or money service	0.67	0.64	0.66	2031
Mortgage	0.90	0.81	0.85	5070
Payday loan, title loan, or personal loan	0.19	0.53	0.28	573
Student loan	0.91	0.76	0.83	6251
Vehicle loan or lease	0.49	0.56	0.52	1761
accuracy			0.58	50968
macro avg	0.53	0.57	0.52	50968
weighted avg	0.65	0.58	0.60	50968

Still not balanced enough.

Results with SVM: adding ***issue***

- Principal components: 20
- Narrative cleaned → 0.98 avg. model accuracy
- Products balanced → 0.96 avg. model accuracy *(less guessing)*
- Removing products that rarely occur → 0.96 avg. model accuracy

Results with SVM: adding *other columns*

- Principal components: 5
- Narrative + issue + …
  - Run 1: ***Sub-issue***
  - Run 2: ***issue*** & ***sub-issue*** *in one column*
  - Run 3: ***State***
  - Run 4: ***ZIP-code***
  - Run 5: ***Company***

Run	min_df	max_df	PC	kernel	accuracy
1	2	0.5	5	linear	0.99
2	2	0.5	5	linear	0.99
3	2	0.5	5	linear	0.98
4	2	0.5	5	linear	0.98
5	2	0.5	5	linear	0.98

Observations:

Thus, minor improvement (+3%) noticeable by adding one more column.

PC of 5 is already enough to yield great results.

Results with RF: ***narrative*** only

- Unbalanced dataset (as an experiment)

Run	PC	accuracy	
1	5	0.57	… without lemmatization
2	30	0.69	
3	100	0.69	
4	200	0.69	
5	30	0.68	… with lemmatization

… lemmatization did not improve model.

- Balanced dataset

Run	PC	accuracy
6	50	0.55
7	100	0.66

… noticeable improvement when doubling PC.

Results with RF: adding ***issue***

- Principal components: 20
- Narrative cleaned → 0.98 avg. model accuracy
- Products balanced → 0.95 avg. model accuracy *(less guessing)*
- **Products balanced a second time → 0.94 avg. model accuracy *(less guessing)***

Confusion matrix (part of)

- Part of the **confusion matrix** for the run where products were balanced twice

(avg. accuracy 0.94)

← 'loan' & 'loan'

Classification Report of the same run

	precision	recall	f1-score	support
Bank account or service	0.99	0.99	0.99	3715
Checking or savings account	0.99	0.99	0.99	4653
Consumer Loan	0.86	0.81	0.84	2544
Credit card	0.98	0.92	0.95	5020
Credit card or prepaid card	0.94	0.97	0.95	4555
Credit reporting	0.99	1.00	0.99	4791
Credit reporting, credit repair services, or other personal consumer reports	0.98	0.87	0.92	5843
Debt collection	0.98	0.96	0.97	5451
Money transfer, virtual currency, or money service	0.93	0.82	0.87	2099
Money transfers	0.37	0.96	0.53	135
Mortgage	0.97	0.96	0.97	4659
Other financial service	0.09	1.00	0.17	8
Payday loan	0.67	0.94	0.78	316
Payday loan, title loan, or personal loan	0.75	0.88	0.81	1332
Prepaid card	0.81	0.98	0.89	289
Student loan	0.97	0.99	0.98	5244
Vehicle loan or lease	0.62	0.80	0.70	1565
Virtual currency	0.00	0.00	0.00	0
accuracy			0.94	52219
macro avg	0.77	0.88	0.80	52219
weighted avg	0.95	0.94	0.94	52219

Results with RF: adding *other columns*

- Principal components: 5
- Narrative + issue + …
  - Run 1: ***Sub-issue***
  - Run 2: ***State***
  - Run 3: ***ZIP-code***
  - Run 4: ***Company***

Run	min_df	max_df	PC	accuracy
1	2	0.5	5	0.99
2	2	0.5	5	0.98
3	2	0.5	5	0.97
4	2	0.5	5	0.97

Observations:

Thus, minor improvement (between 3 & 5%) noticeable by adding one more column.

PC of 5 is already enough to yield great results.

Conclusion

- Possible improvements
  - Try other word embedding technique (e.g. Doc2Vec).
  - More cleaning and data balancing.
  - Use additional methods to measure performance (such as ROC).

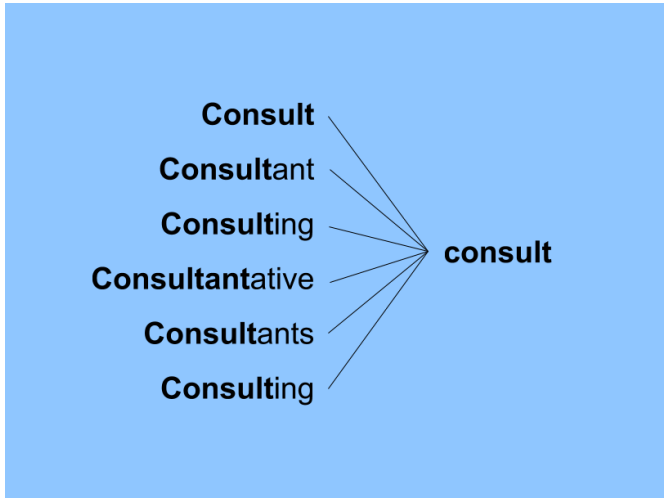
- Observations
  - SVM and RF matched in results with multiple columns, but RF > SVM in *narrative only*.
  - High avg. accuracy when combining narrative & issue.
  - 2-5% avg. accuracy increase when adding a second column, e.g. Sub-issue or Company.
    - *Worth it?*
  - Macro avg. precision could increase if there would be a way to address problems with 'loan' products.

Sprint 2: unsupervised learning

- Finding a different classification
- Exploring new word embedding methods (**Doc2Vec**)
- Unsupervised Learning Algorithms: **K-Means**, **DBSCAN**, **K-Medoids**

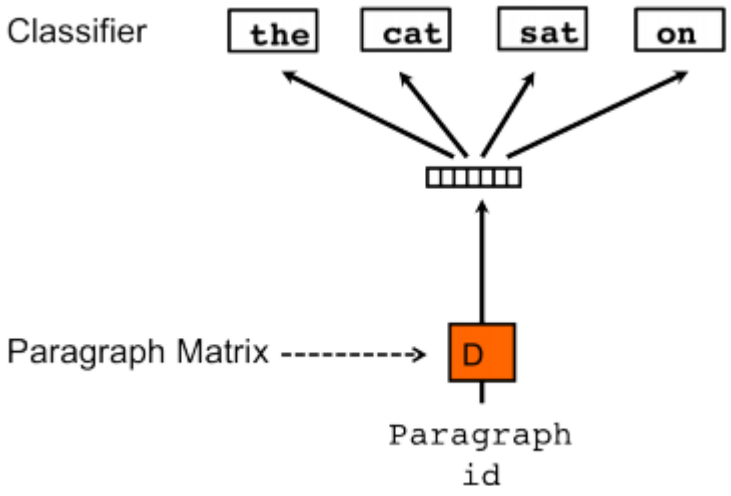
  → Do not use target; clustering algorithms (k clusters)

- Addition cleaning & balancing done, incl. stemming instead of lemmatization



Doc2Vec: Representation of the *narrative*

- Different word embedding: **Doc2Vec** *(DBOW variant)*
- Tagged Document
- Build Vocabulary



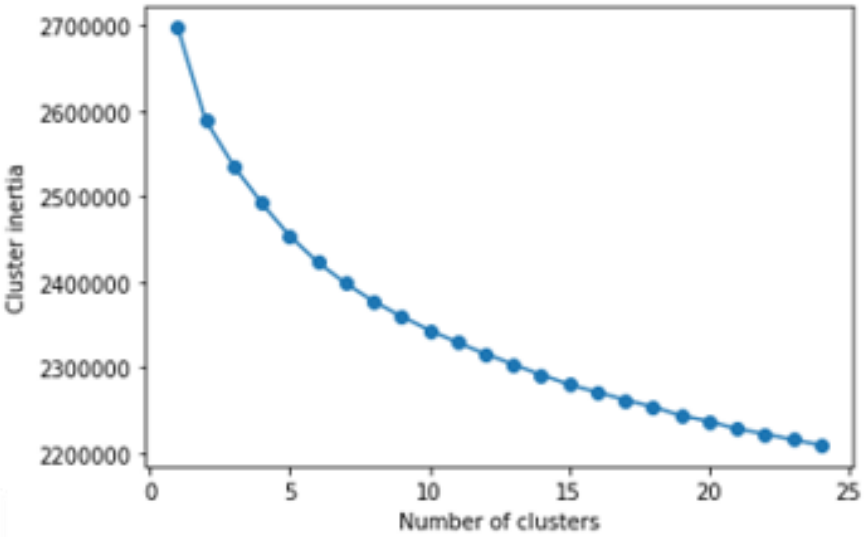
Architecture of PV-DBOW (Mikolov et al., 2014)

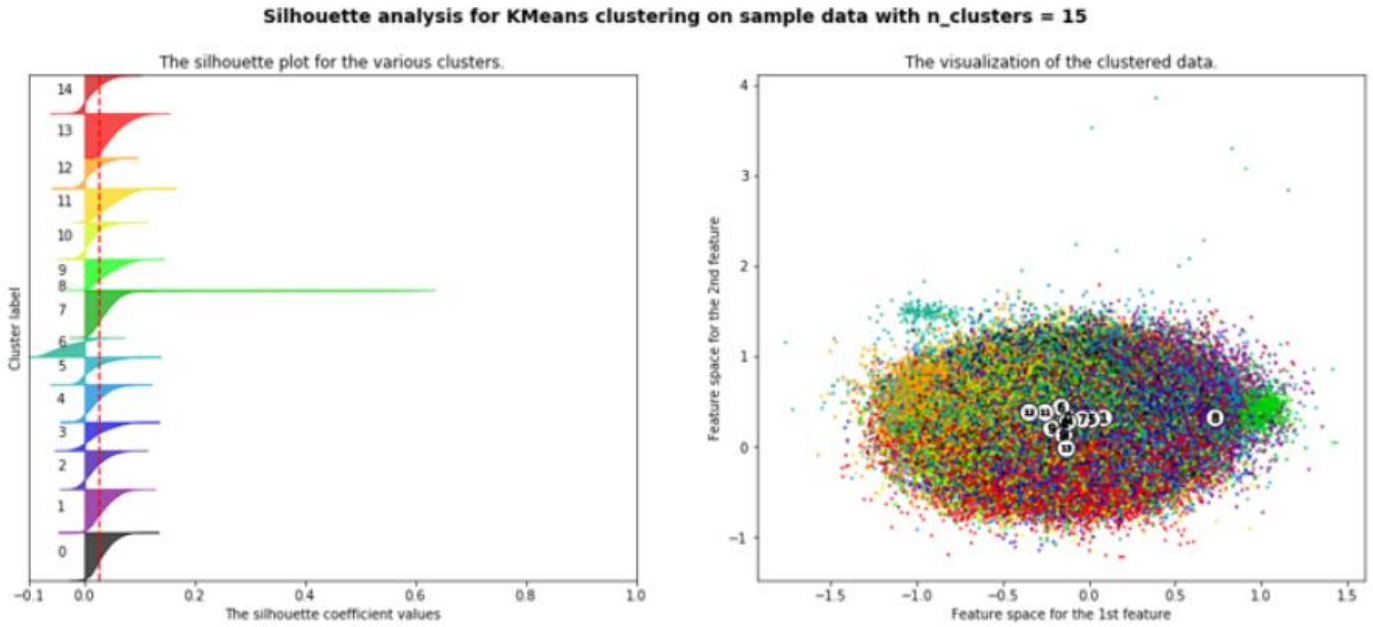
Observations:

Possible improvement: fine-tune hyper parameters for Doc2Vec.
Here: followed the recommendations by Lau and Baldwin.

Results with K-Means

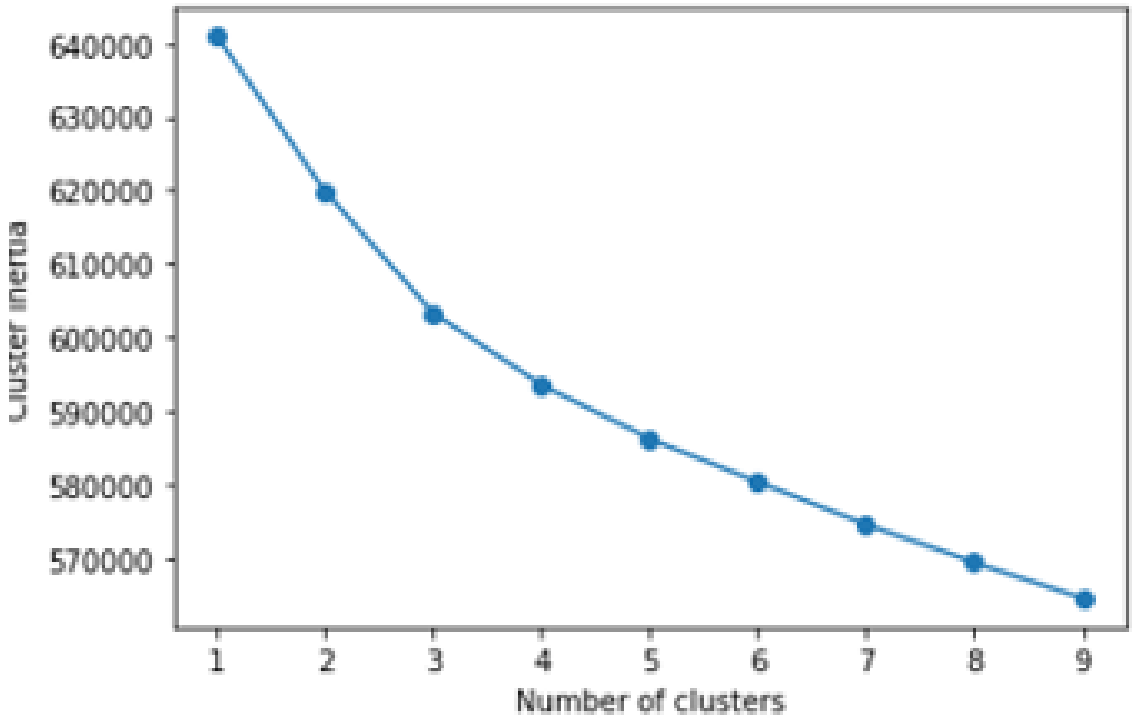
- Unbalanced data set





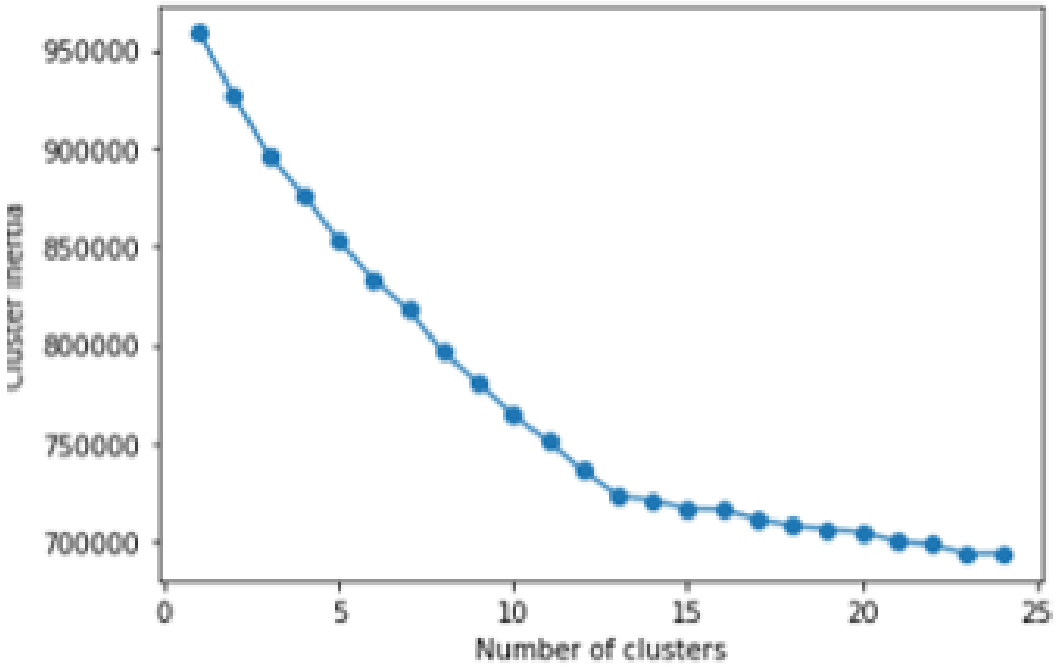
Results with K-Means

- Unbalanced data set
- Balanced data set



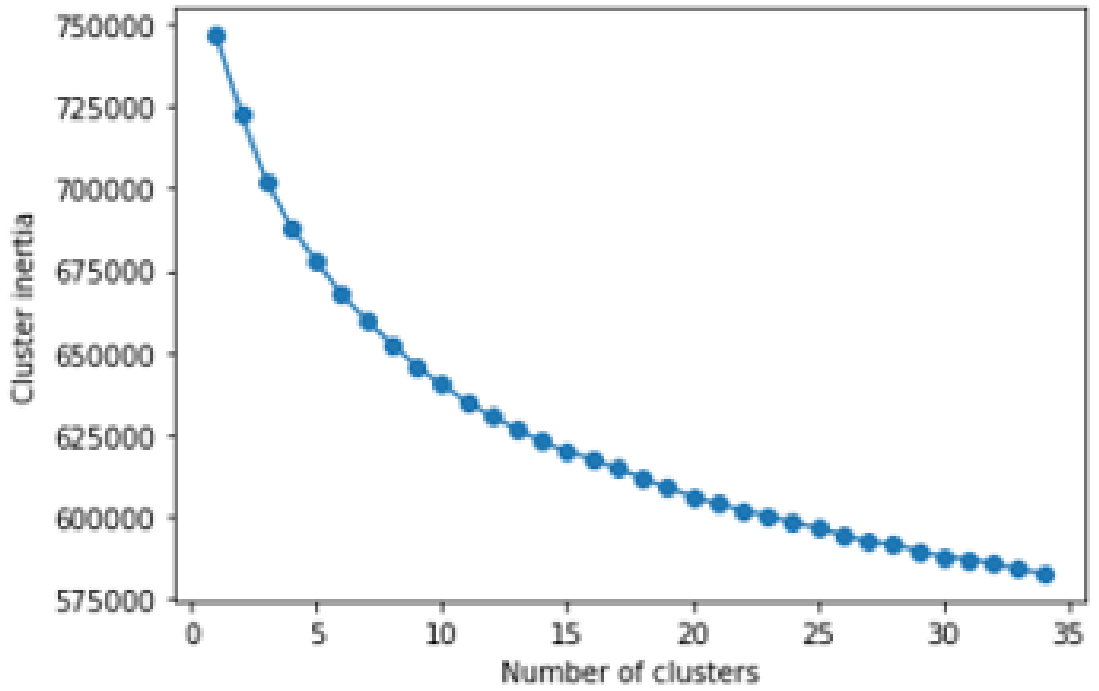
Results with K-Means

- Unbalanced data set
- Balanced data set
  - Runs containing ‘product’



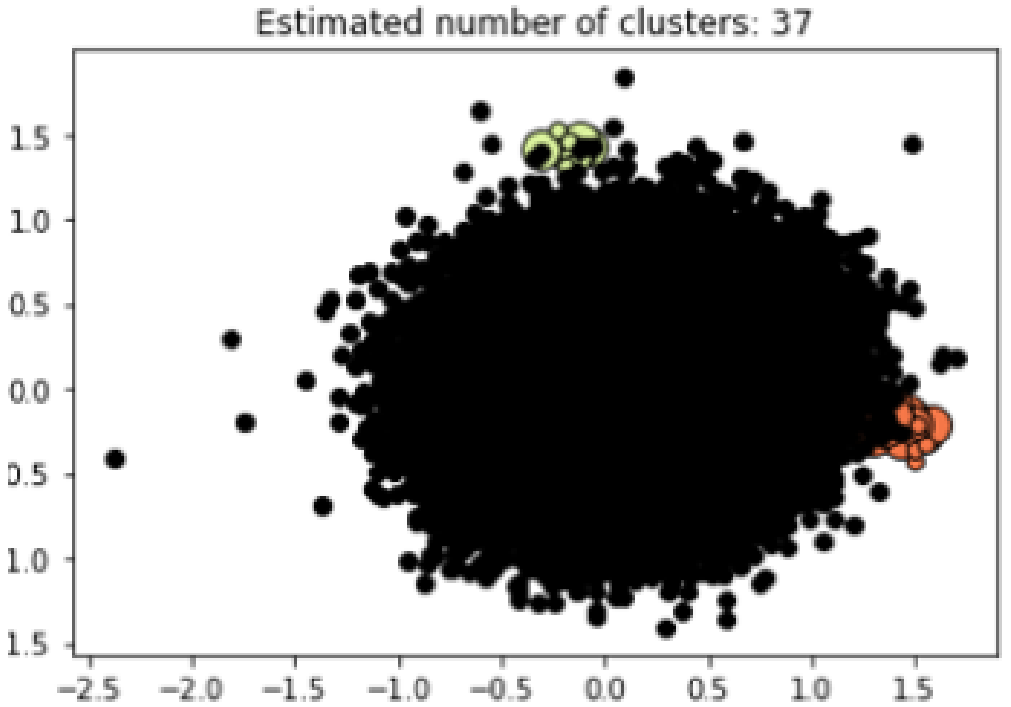
Results with K-Means

- Unbalanced data set
- Balanced data set
  - Runs containing ‘product’
  - Runs not containing ‘product’



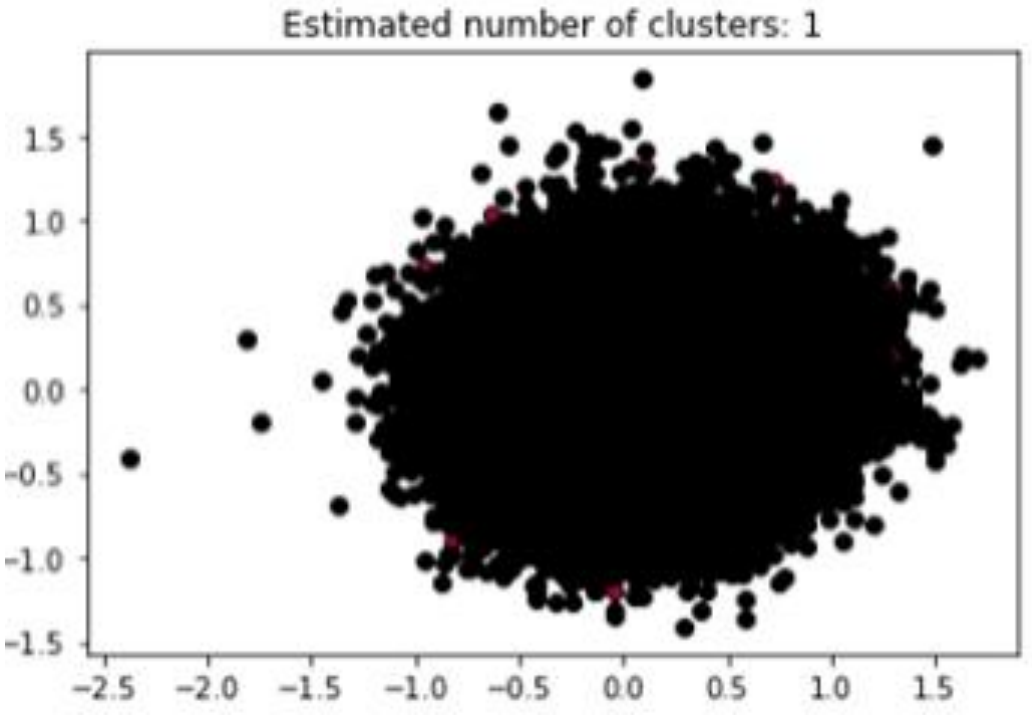
Results with DBSCAN

- Balanced data set: ε = 0.7 min_samples = 100



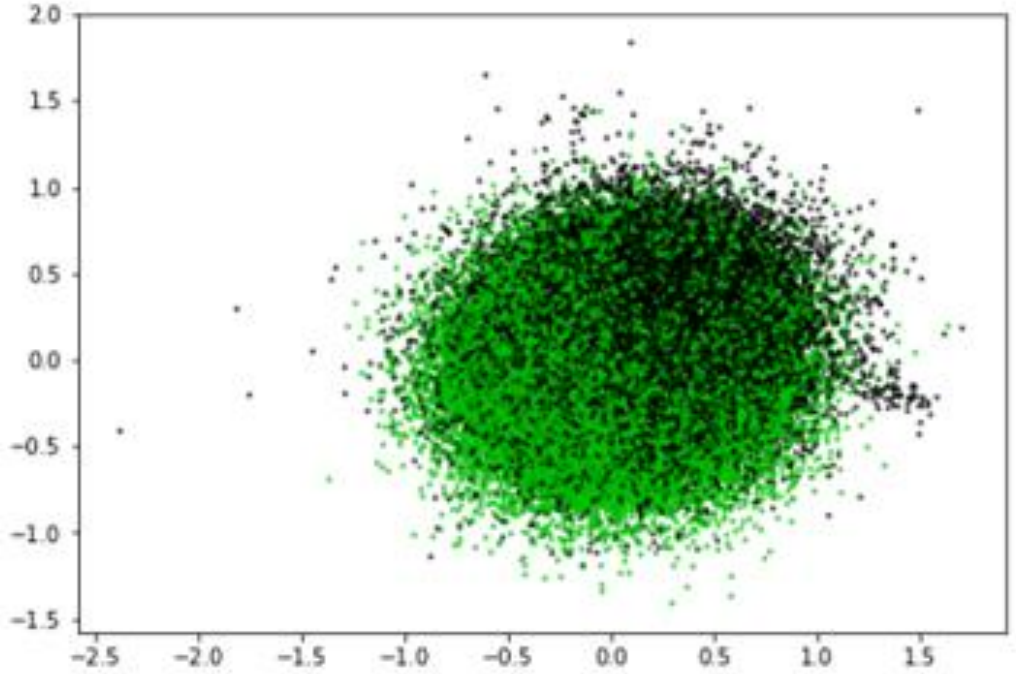
Results with DBSCAN

- Balanced data set: ε = 2 min_samples = 100



Results with K-Medoids

- Balanced data set

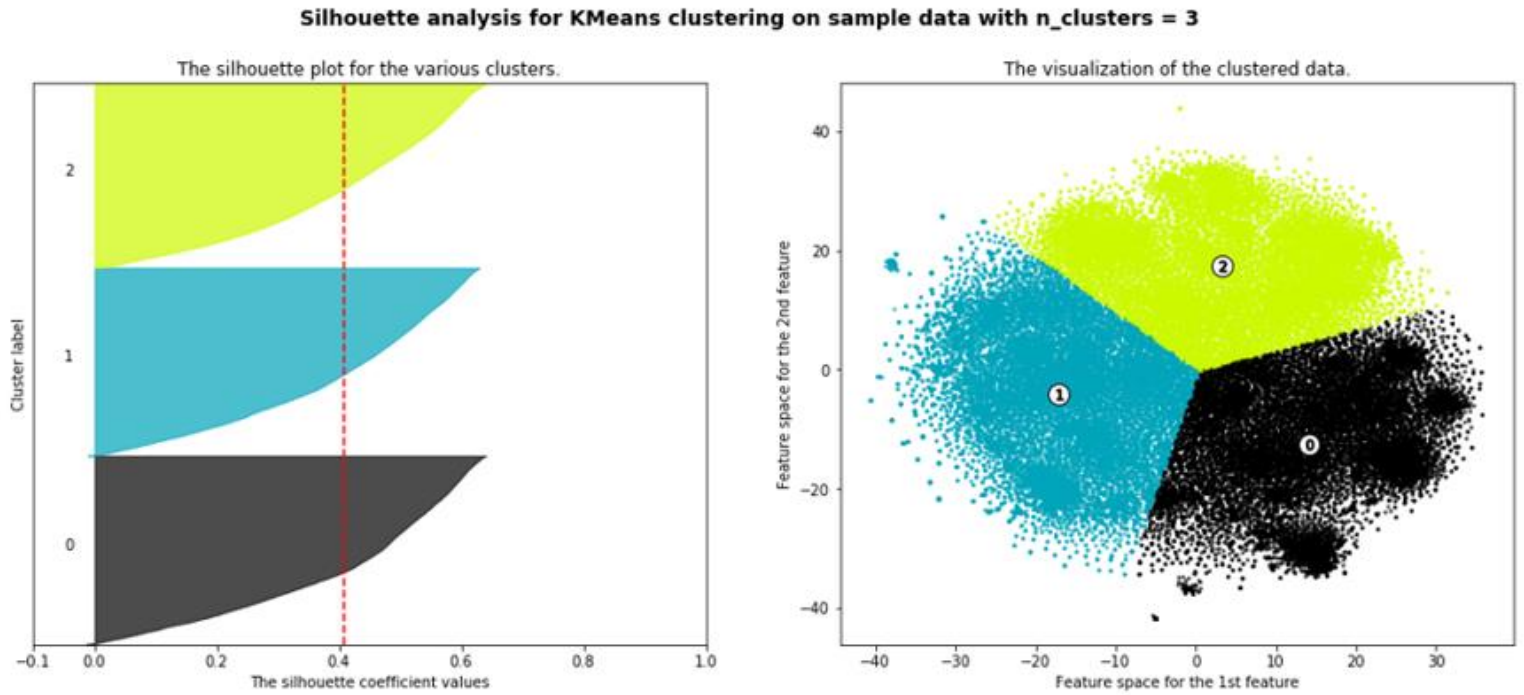


Possible improvements

- Dimensionality Reduction
- Better Visualisation
- Higher k-value
- Show examples of *narratives* in different clusters

K-Means with improvements: T-SNE

- *Narrative* only

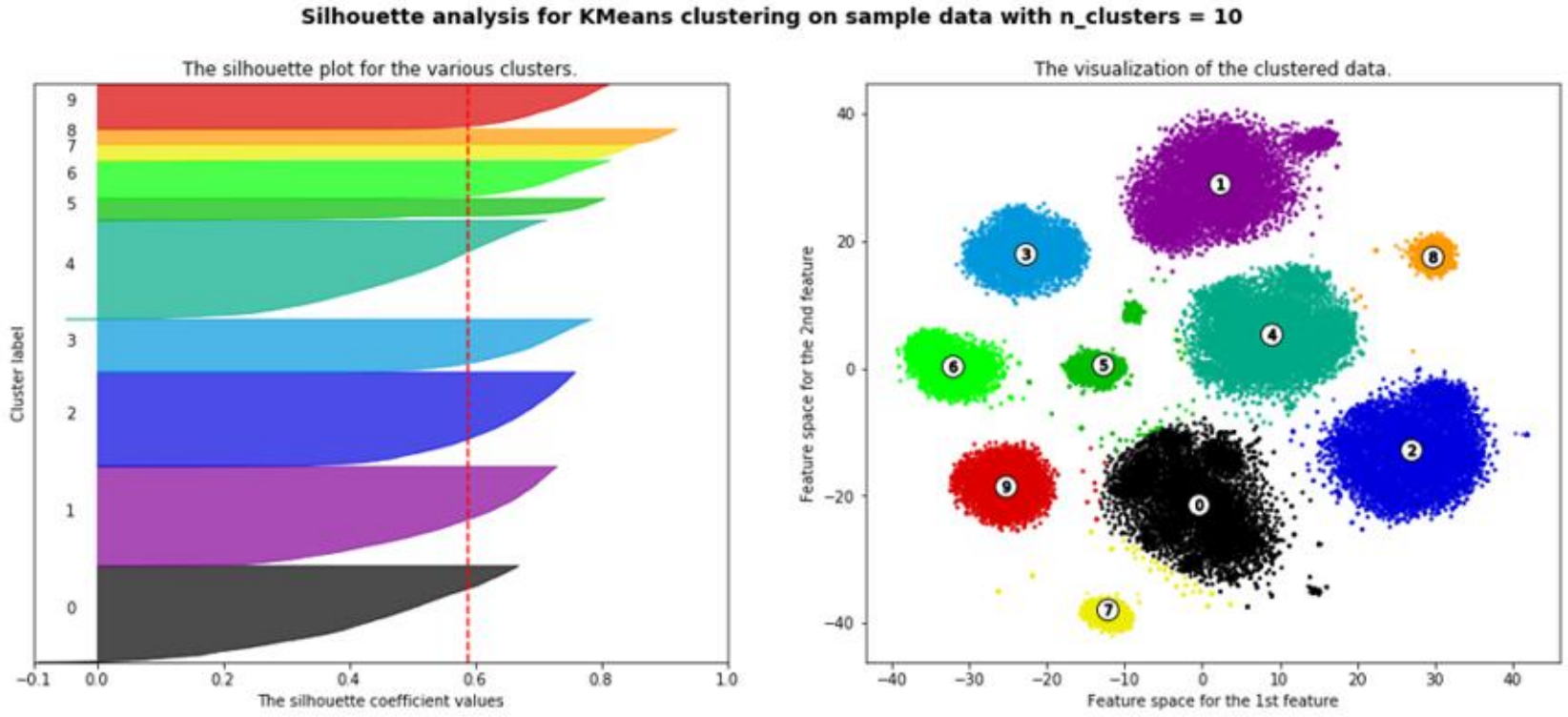


K-Means with improvements: T-SNE

- *Narrative* only

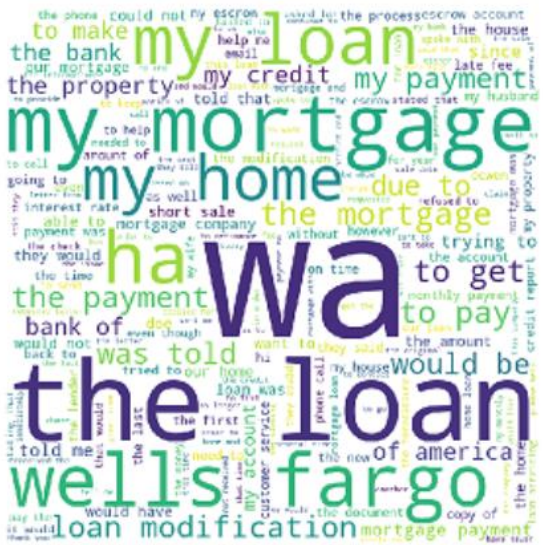
K-Means with improvements: T-SNE

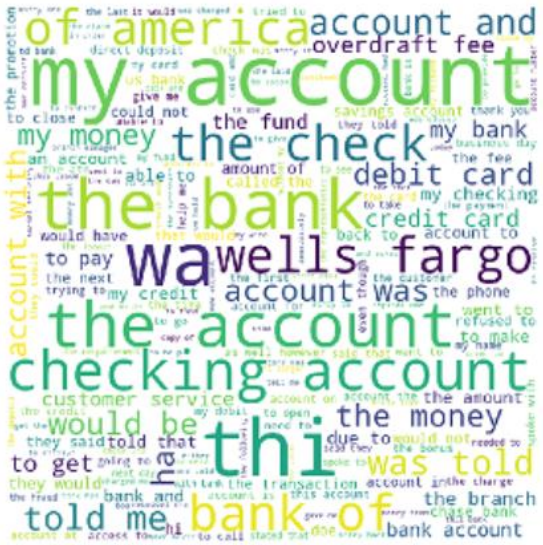
- *Narrative* and *Product*

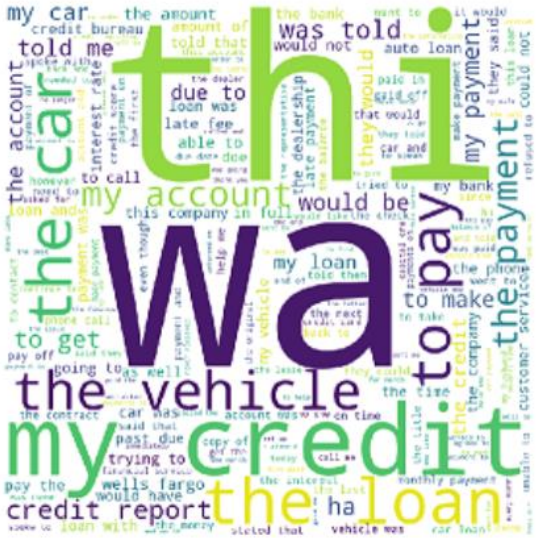


K-Means with improvements: T-SNE

- *Narrative* and *Product*

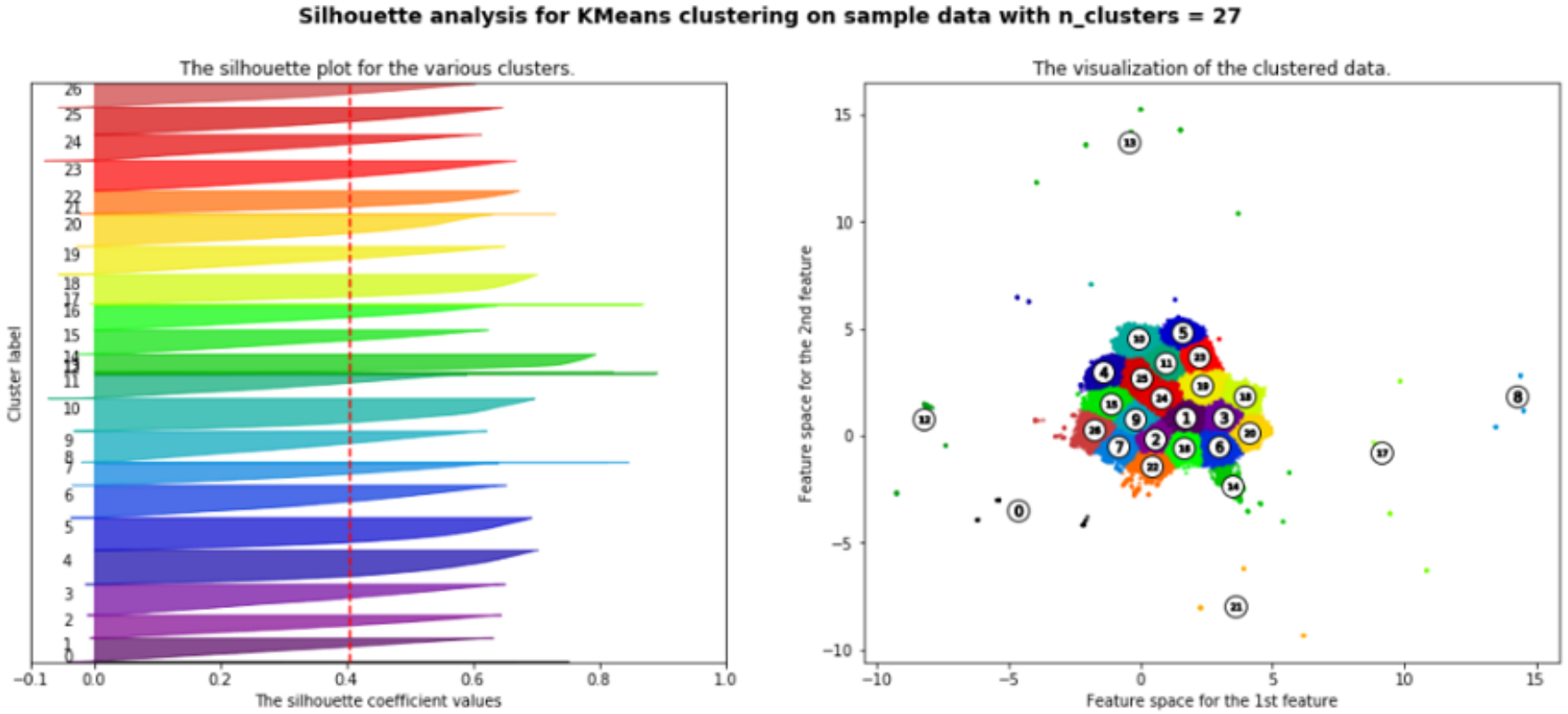






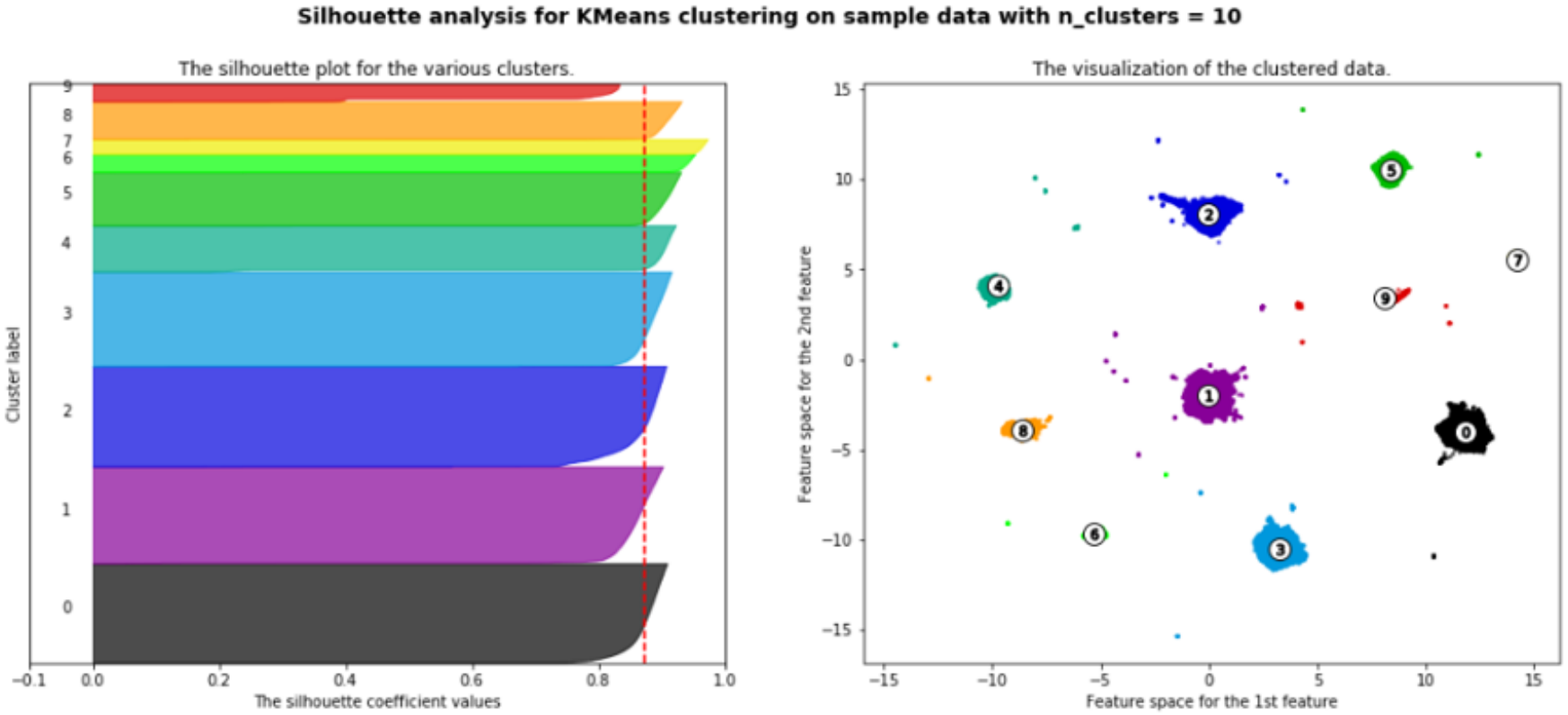
K-Means with improvements: UMAP

- *Narrative* only



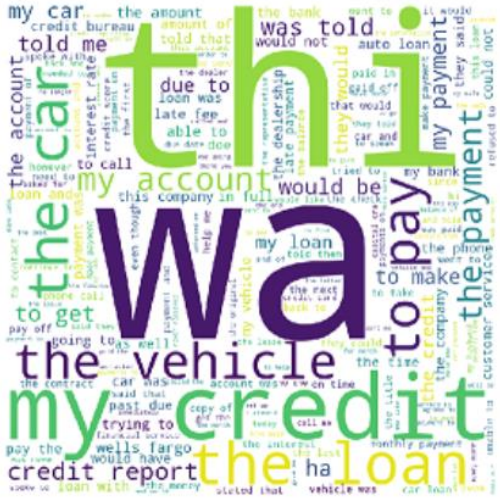
K-Means with improvements: UMAP

- *Narrative* and *Product*



K-Means with improvements: UMAP

- *Narrative* and *Product*

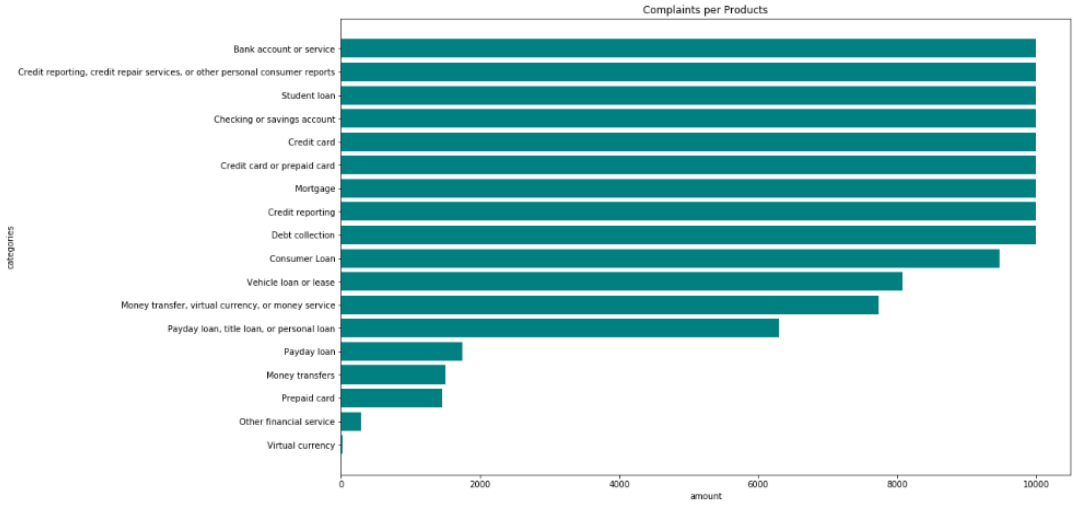


Sprint 3: LDA and BERT

- Additional pre-processing
- Unsupervised Learning Algorithm (Topic Modelling): **LDA**
- Supervised Learning Algorithm (Deep Neural Network): **BERT**

Additional pre-processing

- For LDA
  - Stemming with PorterStemmer instead of Lemmatization
  - Additional balancing

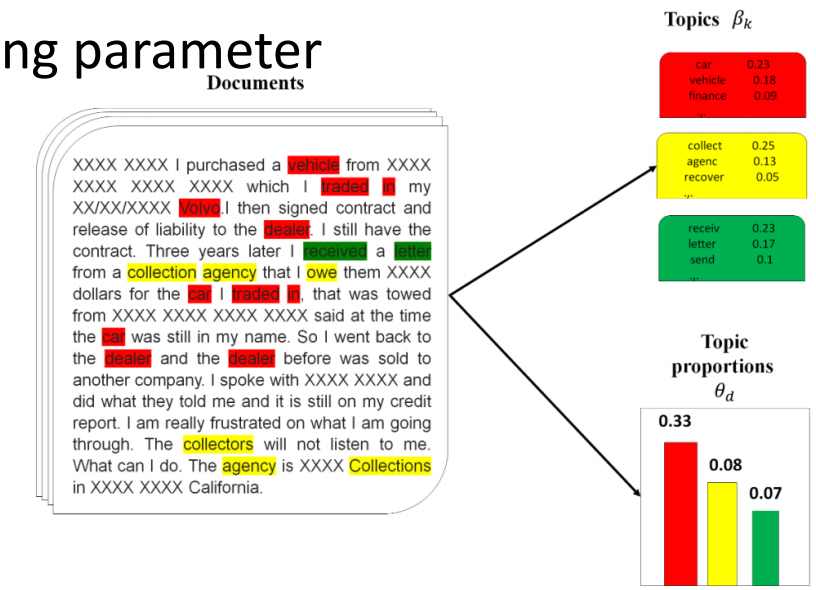


  - Remove additional stop words
    - Common English words;
    - State names;
    - Other frequent words that add little to no value to context.

- Word embedding with Doc2Vec

Latent Dirichlet Allocation (LDA)

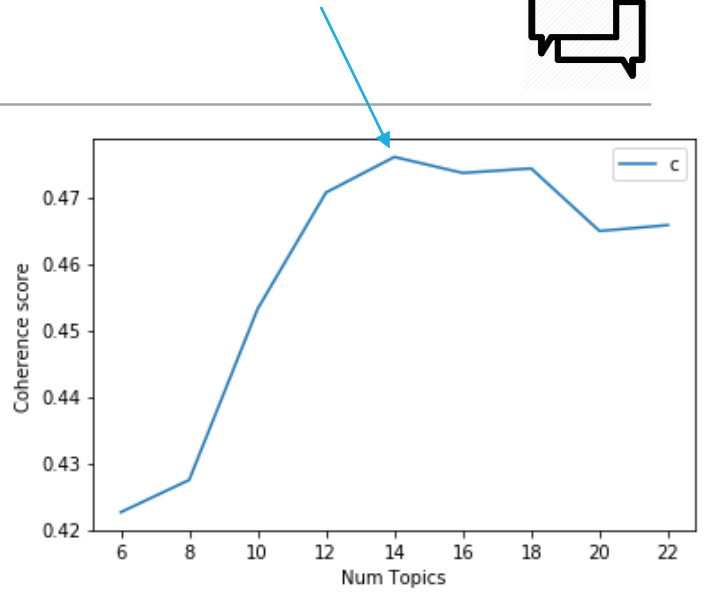
- **Topic modelling algorithm** (unsupervised learning)
- Gensim library
- Two main inputs for the model:
  - Dictionary (id2word)
  - Corpus
- Topic smoothing parameter & Word/term smoothing parameter
  → Blei et al. give suggestions.
- Find optimal number of topics
- 3 versions:
  - balanced data set
  - balanced data set (w/ more stop words removed)
  - Imbalanced data set (as an experiment)



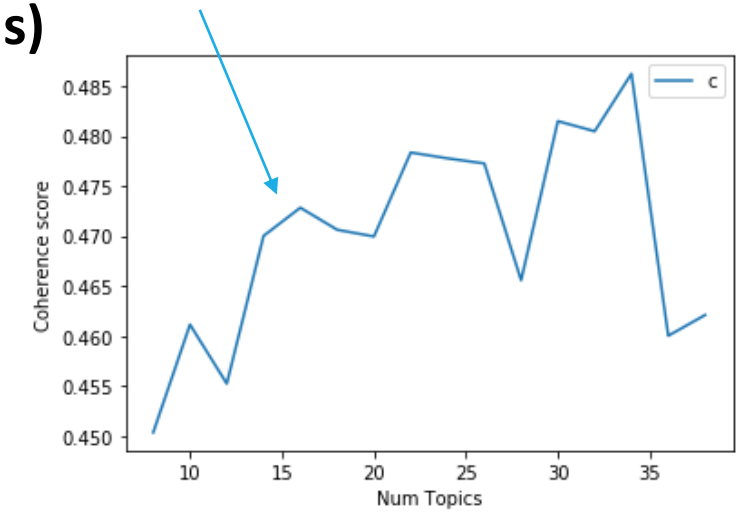
Source: Bastani et al.

Coherence Score

- Coherence Score (balanced version)



- **Coherence Score (balanced version w/ more stop words)**



Observation:

Select topics value (k) that marks end of rapid growth in topic coherence; before flattening.

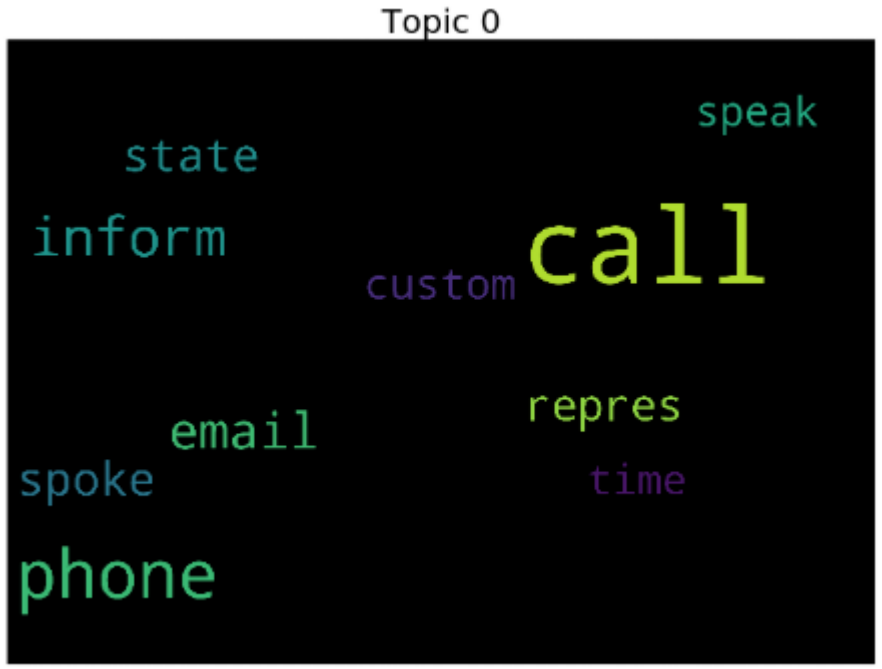
Topics detected by LDA

Field Expert required

```
Start with number of topics: 14
Saving the model...
Computing complexity...
Perplexity:  -6.605748223035092
Computing coherence...
Coherence Score:  0.4700041362723061

Topic: 0
Words: 0.082*"call" + 0.024*"phone" + 0.015*"inform" + 0.013*"email" + 0.011*"spoke" + 0.011*"state" + 0.010*"repres" + 0.009*"custom" + 0.009*"time" + 0.009*"speak"

Topic: 1
Words: 0.054*"mortgag" + 0.034*"loan" + 0.028*"home" + 0.016*"properti" + 0.016*"modif" + 0.014*"tax" + 0.013*"document" + 0.011*"hous" + 0.010*"foreclosur" + 0.010*"escrow"

Topic: 2
Words: 0.104*"credit" + 0.102*"report" + 0.025*"inform" + 0.024*"remov" + 0.023*"disput" + 0.016*"bureau" + 0.016*"debt" + 0.014*"letter" + 0.013*"collect" + 0.013*"file"

Topic: 3
Words: 0.043*"car" + 0.024*"vehicl" + 0.015*"financ" + 0.015*"payment" + 0.011*"loan" + 0.011*"work" + 0.010*"purchas" + 0.010*"credit" + 0.009*"compani" + 0.008*"money"

Topic: 4
Words: 0.122*"card" + 0.059*"credit" + 0.039*"charg" + 0.018*"purchas" + 0.018*"disput" + 0.013*"capit" + 0.012*"close" + 0.010*"use" + 0.010*"servic" + 0.009*"cancel"

Topic: 5
Words: 0.026*"insur" + 0.024*"bill" + 0.021*"letter" + 0.016*"servic" + 0.016*"leas" + 0.014*"return" + 0.013*"financi" + 0.012*"pnc" + 0.011*"request" + 0.010*"provid"

Topic: 6
Words: 0.116*"payment" + 0.050*"fee" + 0.037*"charg" + 0.032*"balanc" + 0.031*"late" + 0.023*"interest" + 0.019*"amount" + 0.018*"statement" + 0.012*"bank" + 0.011*"monthli"

Topic: 7
Words: 0.047*"bank" + 0.044*"money" + 0.033*"fund" + 0.027*"transfer" + 0.025*"transact" + 0.011*"email" + 0.010*"custom" + 0.010*"access" + 0.009*"withdraw" + 0.008*"hold"

Topic: 8
Words: 0.029*"offer" + 0.024*"credit" + 0.022*"rate" + 0.018*"applic" + 0.018*"promot" + 0.017*"open" + 0.016*"appli" + 0.015*"bonu" + 0.014*"requir" + 0.013*"approv"

Topic: 9
Words: 0.071*"loan" + 0.052*"debt" + 0.040*"compani" + 0.030*"collect" + 0.025*"owe" + 0.023*"amount" + 0.015*"interest" + 0.014*"call" + 0.011*"money" + 0.010*"payment"

Topic: 10
Words: 0.022*"consum" + 0.016*"inform" + 0.015*"provid" + 0.013*"law" + 0.011*"complaint" + 0.011*"request" + 0.011*"violat" + 0.010*"act" + 0.009*"document" + 0.008*"actio
n"

Topic: 11
Words: 0.038*"fraud" + 0.029*"bank" + 0.023*"inform" + 0.023*"fraudul" + 0.022*"name" + 0.022*"claim" + 0.019*"address" + 0.018*"report" + 0.018*"secur" + 0.017*"file"

Topic: 12
Words: 0.110*"check" + 0.094*"bank" + 0.034*"deposit" + 0.026*"close" + 0.022*"money" + 0.020*"branch" + 0.019*"cash" + 0.012*"went" + 0.012*"open" + 0.011*"call"

Topic: 13
Words: 0.094*"loan" + 0.040*"payment" + 0.024*"student" + 0.024*"navient" + 0.013*"incom" + 0.013*"plan" + 0.012*"repay" + 0.011*"program" + 0.011*"servic" + 0.010*"school"
**********
```

Word clouds for k = 14

- Example topics:



Communications; customer support

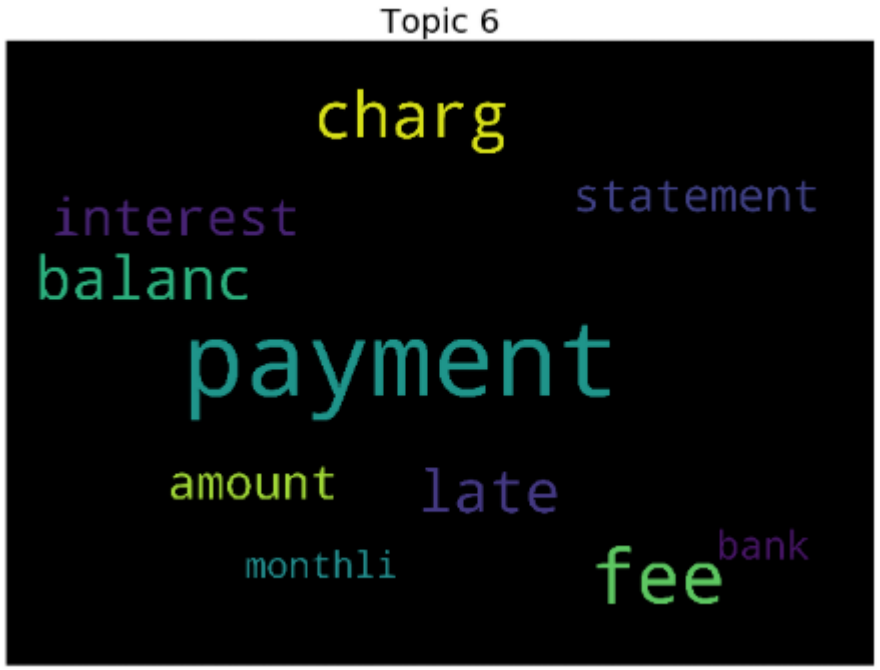
Word clouds for k = 14

- Example topics:

Housing or property / mortgages

Word clouds for k = 14

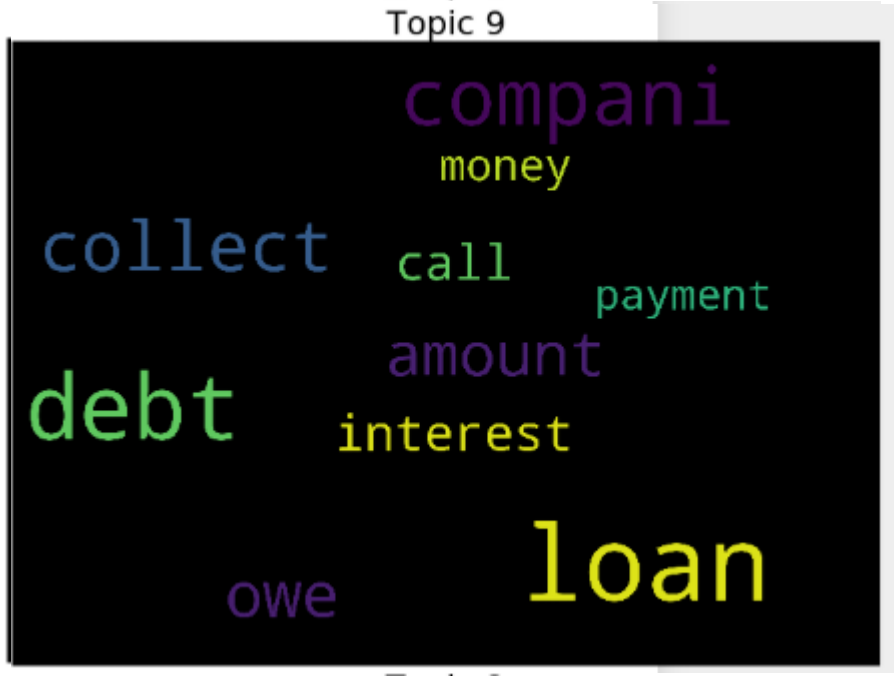
- Example topics:



Bank payments

Word clouds for k = 14

- Example topics:



General loans & debt related

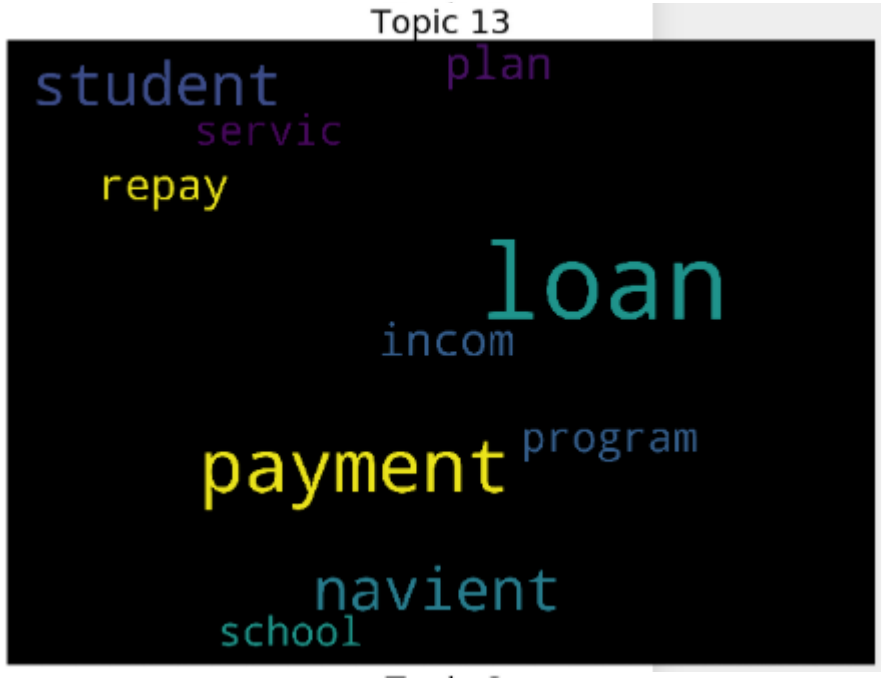
Word clouds for k = 14

- Example topics:

Fraud

Word clouds for k = 14

- Example topics:

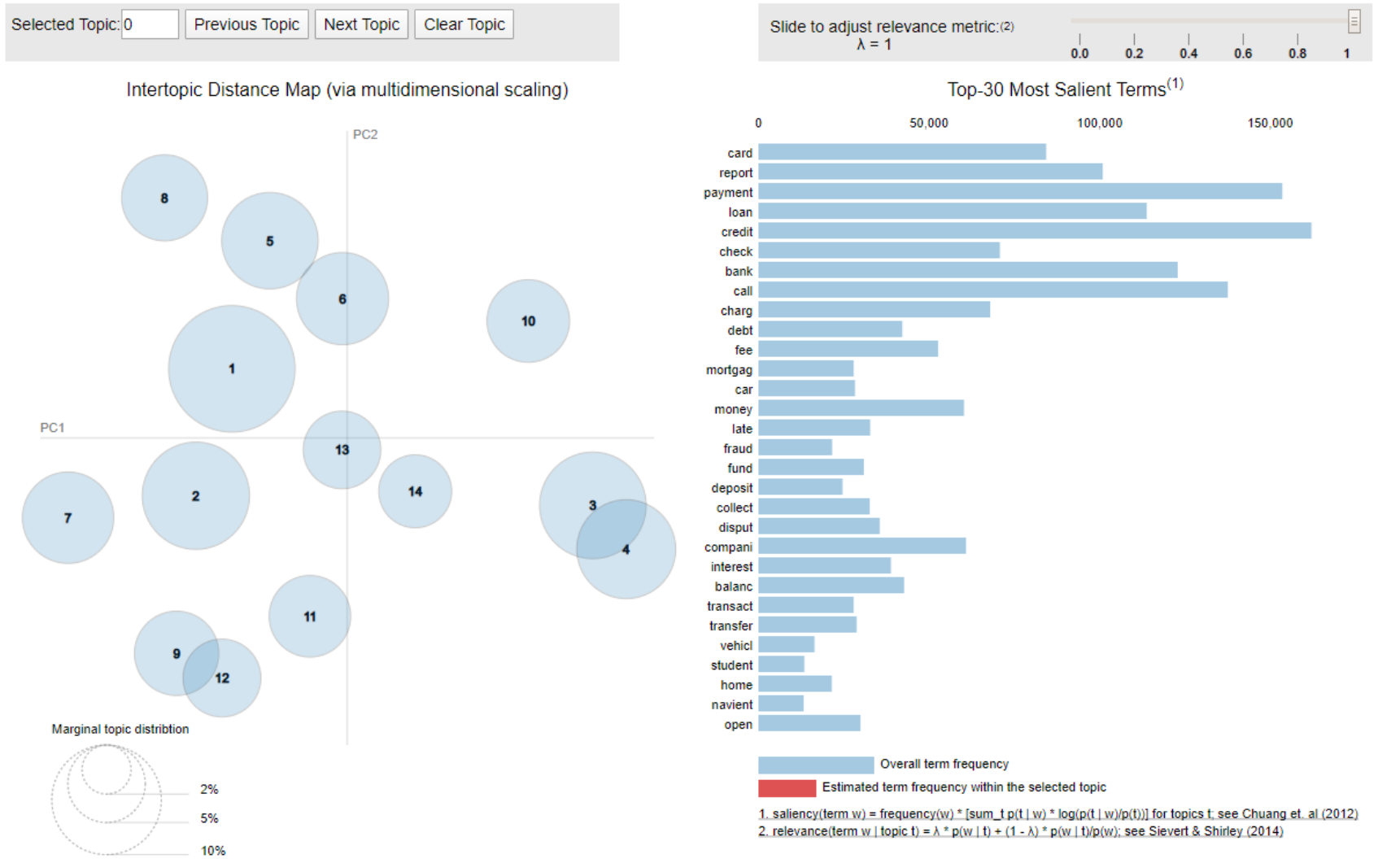


Student loans

Dominant topic & its contribution

	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	6.0	0.3151	payment, fee, charg, balanc, late, interest, a...	they would not let me pay my loan off days bef...
1	1	6.0	0.3248	payment, fee, charg, balanc, late, interest, a...	service finance are liars and are charging me ...
2	2	6.0	0.4851	payment, fee, charg, balanc, late, interest, a...	on i signed a car loan agreement to finance my...
3	3	0.0	0.4641	call, phone, inform, email, spoke, state, repr...	we hired and debt collection to handle collect...
4	4	9.0	0.4798	loan, debt, compani, collect, owe, amount, int...	i borrowed in an financial emergency from offi...
5	5	2.0	0.4708	credit, report, inform, remov, disput, bureau,...	prestige auto has failed to update the past du...
6	6	5.0	0.4181	insur, bill, letter, servic, leas, return, fin...	my leased was repossessed in and gm financial ...
7	7	4.0	0.3586	card, credit, charg, purchas, disput, capit, c...	i filed a complaint directly with paypal for a...
8	8	13.0	0.3877	loan, payment, student, navient, incom, plan, ...	the loans provided by this provider are predat...
9	9	3.0	0.3544	car, vehicl, financ, payment, loan, work, purc...	i have been harrassed for years ive tried repe...

Visualisation using PyLDAvis

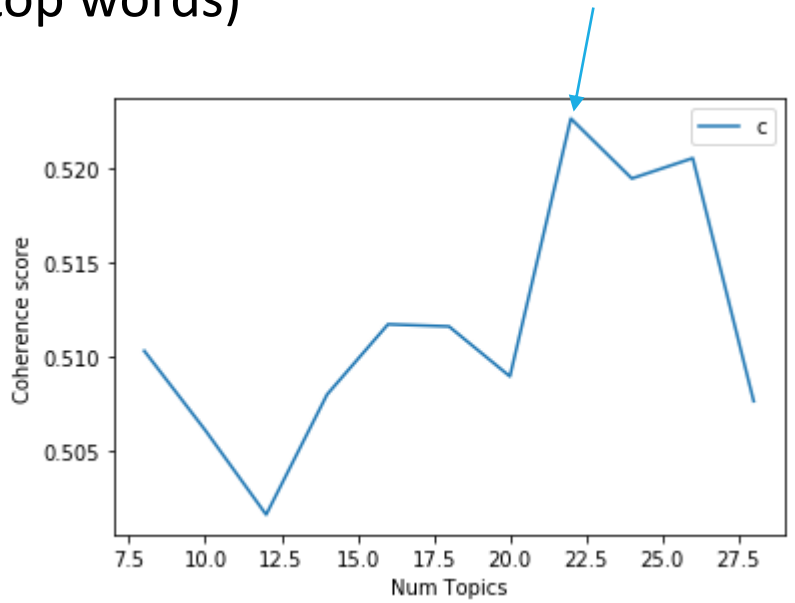


Increase in k:
→ higher chance in topic overlap;
→ small sized bubbles in one region

We desire:
→ balance between k & overlap
→ large, scattered,
non-overlapping bubbles

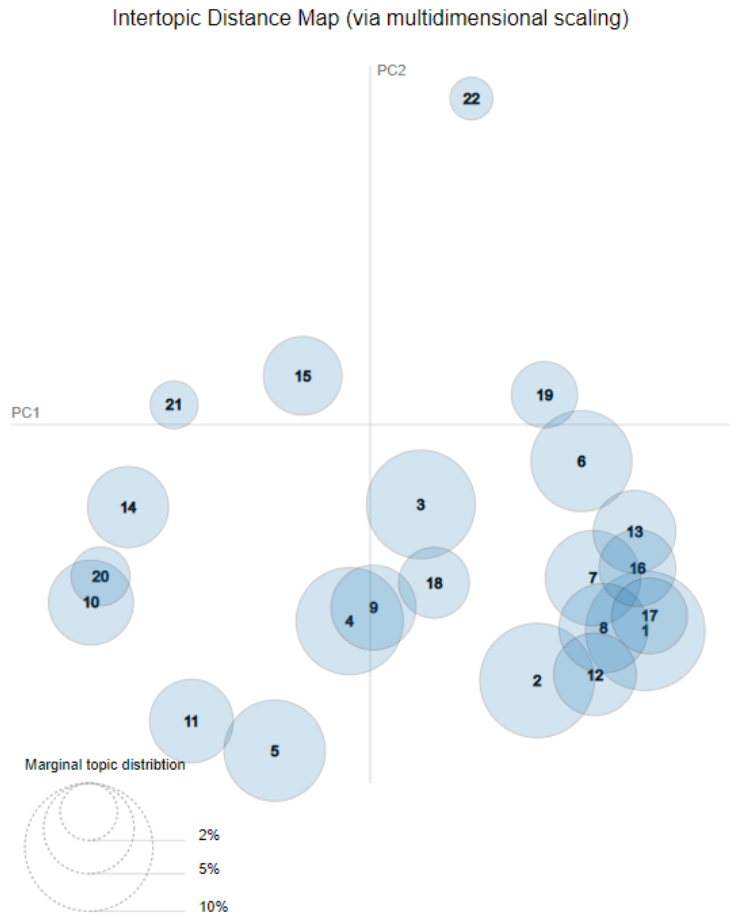
Imbalanced data set

- Coherence Score (imbalanced version w/ more stop words)

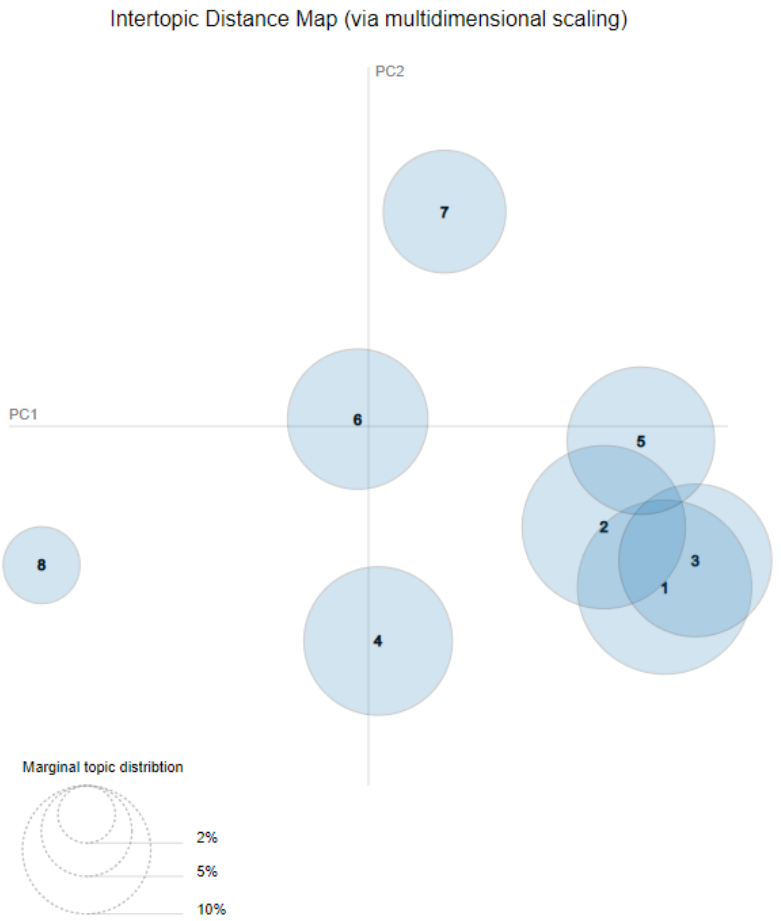


Imbalanced data set

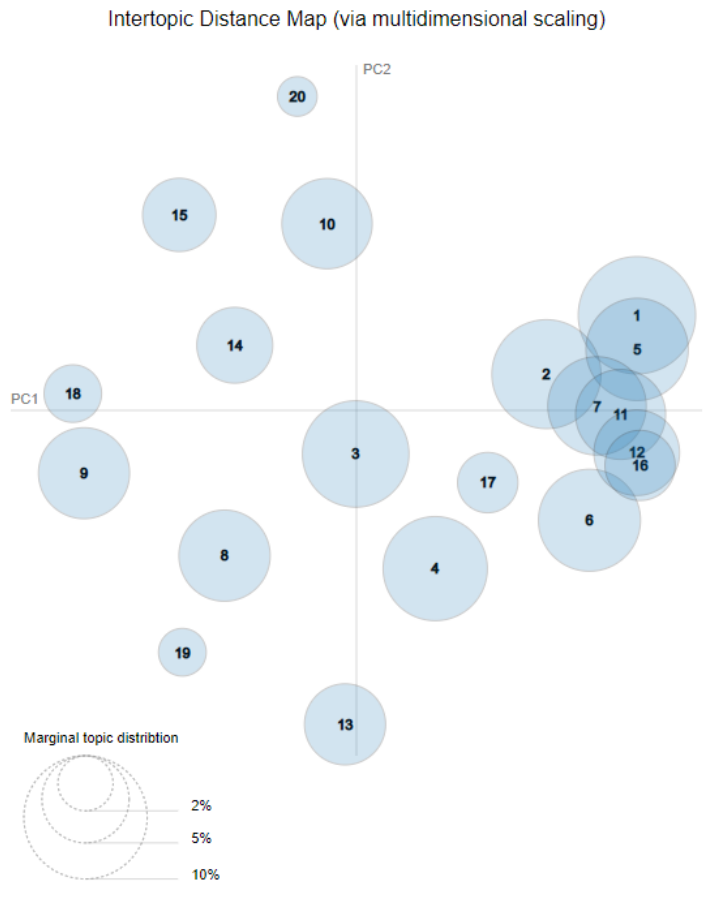
k = 22



k = 8



k = 20



Conclusion

- LDA model, balanced data set, with more stop words removed
  - Improvement over original LDA model with just a balanced data set.
  - k = 14 yields fairly well distributed topics.

- LDA model, imbalanced data set, with more stop words removed
  - Much worse than previous two.

Future work

- Remove even more stop words?
- Attempt t-SNE
- Tweak certain parameters of the model, such as alpha & eta, iterations, passes
- Use different scoring measure for the model (e.g. U mass instead of c_v)
- Use ‘Mallet’ as a wrapper:
  - Gensim provides wrapper to implement Mallet’s LDA
  - Could potentially increase c_v

BERT

- **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (Devlin et al., 2019)
- Deep neural network based technique for NLP
- Word embeddings
- Pre-trained on:
  - Roughly 10,000 books
  - English Wikipedia

- Fine-tune this model → Additional output layer

BERT

- Language model can learn from context based on surrounding words

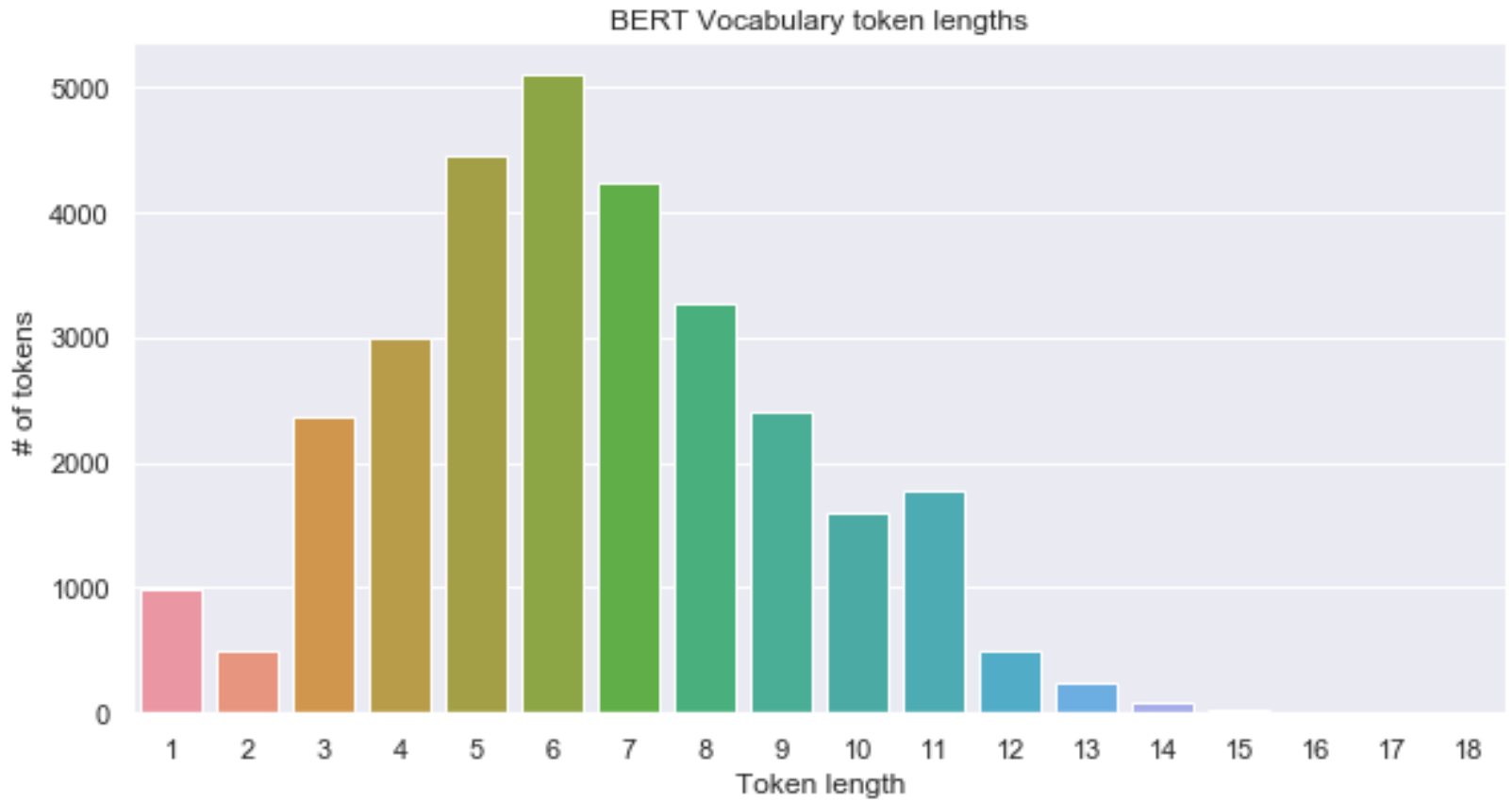
  → Contextual model
- Large & **Base** (~ 30K ‘words’)
- We can use BERT for:
  - **Classification** (supervised learning)
  - Named Entity Recognition
  - Part of Speech Tagging
  - Question Answering
- **Hugging Face** implementation of BERT using **PyTorch**

- **Google Colab**

BERT's vocabulary

- Distribution of word token length

```
Text(0, 0.5, '# of tokens')
```



BERT's vocabulary

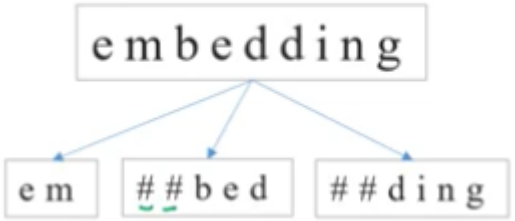
- Misspellings? → It would not appear so.
- Contractions? → No. BERT's tokenizer separates words that contain symbols/punctuation
- Names? → Use Gutenberg's list of names

```
There are 3869 in the BERT vocabulary.
```

- Numbers? Yes, although half of them appear to be dates.

```
The BERT vocab includes 390 out of 521 dates from 1500 - 20201.
This is roughly 44 % of the numbers.
```

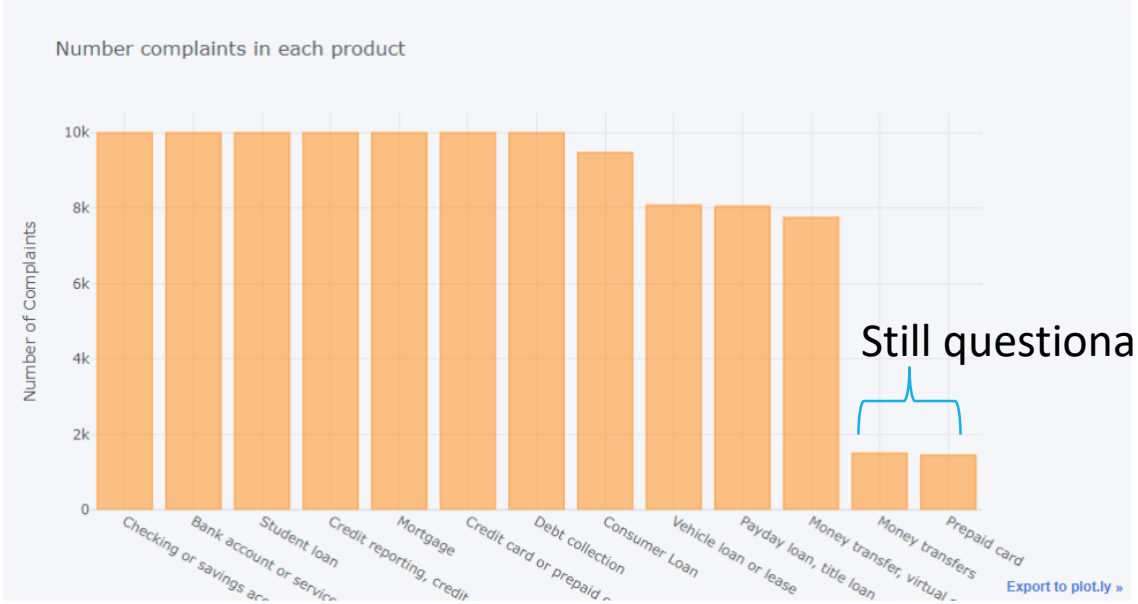
- Symbols? Yes.
- New words?



Fine-tuning BERT

- BERT has its own Tokenizer due to special tags, such as [CLS] and [SEP].

- Additional pre-processing: product consolidation & more balancing
  - E.g. consolidate “Virtual currency” into “Money transfer, virtual currency, or money service.”
  - 18 products → 13 products



Fine-tuning BERT

- BERT has its own Tokenizer due to special tags, such as [CLS] and [SEP].

```
Original Text:  they would not let me pay my loan off days before the next payment plus there were substantial fees that were n
ot made aware to me or misleading not only can you not understand the people in customer service but they are very misleading w
as told i needed hours minimum to pay full loan off before next payment was taken directly out of my account i called two days
before that and they would not let me pay my loan off until a payment was taken out this way they could build up more interest
and make more money since none of the payment goes to principal i dont know how they get away with this predators picking on th
e poor is whats happening and just because its a native american tribe the government cant do anything about it these places ne
ed to be shut down and people need to be more educated about these establishments including myself
```

No special tokens, because this is a fully cleaned complaint, i.e. without punctuation

```
Tokenized Text:  ['they', 'would', 'not', 'let', 'me', 'pay', 'my', 'loan', 'off', 'days', 'before', 'the', 'next', 'payment',
'plus', 'there', 'were', 'substantial', 'fees', 'that', 'were', 'not', 'made', 'aware', 'to', 'me', 'or', 'misleading', 'not',
'only', 'can', 'you', 'not', 'understand', 'the', 'people', 'in', 'customer', 'service', 'but', 'they', 'are', 'very', 'mislead
ing', 'was', 'told', 'i', 'needed', 'hours', 'minimum', 'to', 'pay', 'full', 'loan', 'off', 'before', 'next', 'payment', 'was',
'taken', 'directly', 'out', 'of', 'my', 'account', 'i', 'called', 'two', 'days', 'before', 'that', 'and', 'they', 'would', 'no
t', 'let', 'me', 'pay', 'my', 'loan', 'off', 'until', 'a', 'payment', 'was', 'taken', 'out', 'this', 'way', 'they', 'could', 'b
uild', 'up', 'more', 'interest', 'and', 'make', 'more', 'money', 'since', 'none', 'of', 'the', 'payment', 'goes', 'to', 'princi
pal', 'i', 'don', '##t', 'know', 'how', 'they', 'get', 'away', 'with', 'this', 'predators', 'picking', 'on', 'the', 'poor', 'i
s', 'what', '##s', 'happening', 'and', 'just', 'because', 'its', 'a', 'native', 'american', 'tribe', 'the', 'government', 'ca
n', '##t', 'do', 'anything', 'about', 'it', 'these', 'places', 'need', 'to', 'be', 'shut', 'down', 'and', 'people', 'need', 't
o', 'be', 'more', 'educated', 'about', 'these', 'establishments', 'including', 'myself']
```

```
Token IDs:  [2027, 2052, 2025, 2292, 2033, 3477, 2026, 5414, 2125, 2420, 2077, 1996, 2279, 7909, 4606, 2045, 2020, 6937, 9883,
2008, 2020, 2025, 2081, 5204, 2000, 2033, 2030, 22369, 2025, 2069, 2064, 2017, 2025, 3305, 1996, 2111, 1999, 8013, 2326, 2021,
2027, 2024, 2200, 22369, 2001, 2409, 1045, 2734, 2847, 6263, 2000, 3477, 2440, 5414, 2125, 2077, 2279, 7909, 2001, 2579, 3495,
2041, 1997, 2026, 4070, 1045, 2170, 2048, 2420, 2077, 2008, 1998, 2027, 2052, 2025, 2292, 2033, 3477, 2026, 5414, 2125, 2127, 1
037, 7909, 2001, 2579, 2041, 2023, 2126, 2027, 2071, 3857, 2039, 2062, 3037, 1998, 2191, 2062, 2769, 2144, 3904, 1997, 1996, 79
09, 3632, 2000, 4054, 1045, 2123, 2102, 2113, 2129, 2027, 2131, 2185, 2007, 2023, 12630, 8130, 2006, 1996, 3532, 2003, 2054, 20
15, 6230, 1998, 2074, 2138, 2049, 1037, 3128, 2137, 5917, 1996, 2231, 2064, 2102, 2079, 2505, 2055, 2009, 2122, 3182, 2342, 200
0, 2022, 3844, 2091, 1998, 2111, 2342, 2000, 2022, 2062, 5161, 2055, 2122, 17228, 2164, 2870]
```

Fine-tuning BERT

- Additional pre-processing: data set cleaning → 3 versions

  - **Version 1: Fully cleaned** (lowercasing, removing URLs, removing censored words (XXXX), removing dates, removing all symbols, removing numbers, normalising spaces)

  - **Version 2: Nearly fully cleaned** (lowercasing, removing URLs, removing censored words (XXXX), removing symbols except for punctuation, removing decimal point in number, normalising spaces)

  - **Version 3: Partially cleaned** (basic cleaning: lowercasing, removing URLs, removing newlines & tabs, normalising spaces)

Original ***narrative***

'On XX/XX/XXXX I signed a car loan agreement to finance my 2008 XXXX XXXX XXXX car. \nThe loan amount was {$5400.00}, that is the amount that was financed. The interest was 8.54 %. The total amount to be paid back was {$6500.00}, in 48 months payments. The amount per month was {$130.00}. The payments began on XX/XX/XXXX and were to end on XX/XX/XXXX. Fortunately I made some payments that were much more than the monthly required amount of {$130.00}. That means that the final payment is now to be well before XX/XX/XXXX. In fact I did my last payment in XX/XX/XXXX. In addition I have mailed them a final payment of {$180.00} ({$50.00} deferral fee and {$130.00} for the payment I skipped XX/XX/XXXX). \nDuring the term of loan I received no communication, no monthly account statement and no notice of late payments. Also I have been making my payment on time every month with the exception of the skipped payment in XX/XX/XXXX. \nI have called the bank to let them know I moved from the address that was on the loan agreement. Then called the bank again to let them know I had moved out of state. Then a year later when I moved again to a different address. All 3 times I gave them my new address and the individuals I spoke with each time, stated to me that the address was updated. \nAccording to the instructions in the payment book provided to me at the beginning of the loan it is stated that " we will send you a statement for your final payment amount in advance of your due date \'\'. The due date is the XXXX

Version 1: fully cleaned

'on i signed a car loan agreement to finance my car the loan amount was that is the amount that was financed the interest was t
he total amount to be paid back was in months payments the amount per month was the payments began on and were to end on fortun
ately i made some payments that were much more than the monthly required amount of that means that the final payment is now to
be well before in fact i did my last payment in in addition i have mailed them a final payment of deferral fee and for the paym
ent i skipped during the term of loan i received no communication no monthly account statement and no notice of late payments a
lso i have been making my payment on time every month with the exception of the skipped payment in i have called the bank to le
t them know i moved from the address that was on the loan agreement then called the bank again to let them know i had moved out
of state then a year later when i moved again to a different address all times i gave them my new address and the individuals i
spoke with each time stated to me that the address was updated according to the instructions in the payment book provided to me

Version 2: nearly fully cleaned

"on i signed a car loan agreement to finance my car. the loan amount was . , that is the amount that was financed. the interest was . . the total amount to be paid back was . , in months payments. the amount per month was . . the payments began on and were to end on . fortunately i made some payments that were much more than the monthly required amount of . . that means that the final payment is now to be well before . in fact i did my last payment in . in addition i have mailed them a final payment of . . deferral fee and . for the payment i skipped . during the term of loan i received no communication, no monthly account statement and no notice of late payments. also i have been making my payment on time every month with the exception of the skipped payment in . i have called the bank to let them know i moved from the address that was on the loan agreement. then called the bank again to let them know i had moved out of state. then a year later when i moved again to a different address. all times i gave them my new address and the individuals i spoke with each time, stated to me that the address was updated. according to the i

Version 3: partially cleaned

'on xx/xx/xxxx i signed a car loan agreement to finance my xxxx xxxx xxxx car. the loan amount was {$. }, that is the amount t
hat was financed. the interest was . %. the total amount to be paid back was {$. }, in months payments. the amount per month w
as {$. }. the payments began on xx/xx/xxxx and were to end on xx/xx/xxxx. fortunately i made some payments that were much more
than the monthly required amount of {$. }. that means that the final payment is now to be well before xx/xx/xxxx. in fact i di
d my last payment in xx/xx/xxxx. in addition i have mailed them a final payment of {$. } ({$. } deferral fee and {$. } for
the payment i skipped xx/xx/xxxx). during the term of loan i received no communication, no monthly account statement and no no
tice of late payments. also i have been making my payment on time every month with the exception of the skipped payment in xx/x
x/xxxx. i have called the bank to let them know i moved from the address that was on the loan agreement. then called the bank a

Too many words

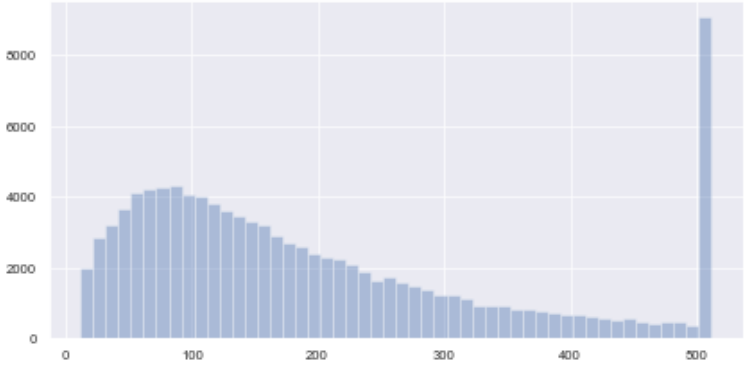
- BERT: max sequence length (i.e. max. tokens or words per complaints) = 512
  - Sequence length for all complaints must be the same. → Use padding [PAD]
  - Use **chunking** or **summerization** if you want > 512

- Idea: complaint **conveys sufficient information** in set sequence length

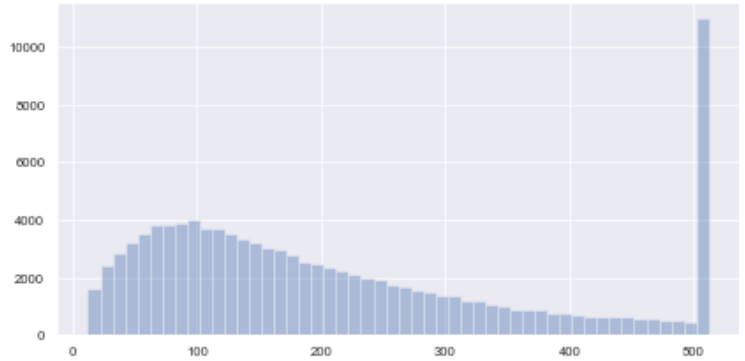
- Truncation strategy:
  - Tokenize text
  - Encode
  - Truncate to desired length

Too many words

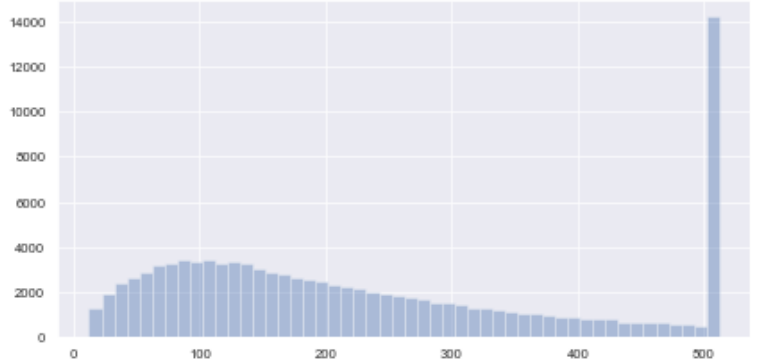
Version 1: fully cleaned



Version 2: nearly fully cleaned



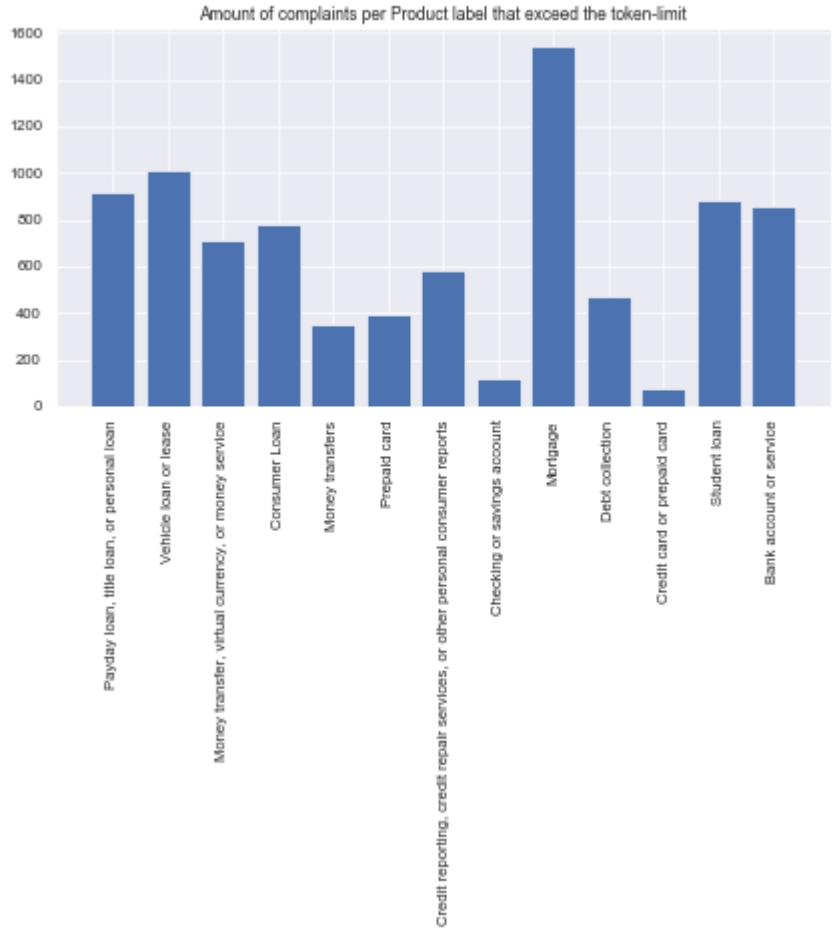
Version 3: partially cleaned



Visually see how many complaints are affected by truncation.

Too many words

- Affected complaints per product
- Version 1 data set
- Most affected: **mortgage**



Attention masks

- Tell BERT which tokens are actual words vs. padding ([PAD], represented as 0)

```
256
[101, 2027, 2052, 2025, 2292, 2033, 3477, 2026, 5414, 2125, 1016, 2420, 2077, 1996, 2279, 7909, 4606, 2045, 2020, 6937, 9883, 2
008, 2020, 2025, 2081, 5204, 2000, 2033, 2030, 22369, 1012, 2025, 2069, 2064, 2017, 2025, 3305, 1996, 2111, 1999, 8013, 2326, 2
021, 2027, 2024, 2200, 22369, 1012, 2001, 2409, 1045, 2734, 2484, 2847, 6263, 2000, 3477, 2440, 5414, 2125, 2077, 2279, 7909, 2
001, 2579, 3495, 2041, 1997, 2026, 4070, 1012, 1045, 2170, 2048, 2420, 2077, 2008, 1998, 2027, 2052, 2025, 2292, 2033, 3477, 20
26, 5414, 2125, 2127, 1037, 7909, 2001, 2579, 2041, 1012, 2023, 2126, 2027, 2071, 3857, 2039, 2062, 3037, 1998, 2191, 2062, 276
9, 2144, 3904, 1997, 1996, 7909, 3632, 2000, 4054, 1012, 1045, 2123, 2102, 2113, 2129, 2027, 2131, 2185, 2007, 2023, 1010, 1263
0, 8130, 2006, 1996, 3532, 2003, 2054, 2015, 6230, 1998, 2074, 2138, 2049, 1037, 3128, 2137, 5917, 1996, 2231, 2064, 2102, 207
9, 2505, 2055, 2009, 1012, 2122, 3182, 2342, 2000, 2022, 3844, 2091, 1998, 2111, 2342, 2000, 2022, 2062, 5161, 2055, 2122, 1722
8, 1010, 2164, 2870, 1012, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0]
```

Data set split

- 80% training set, 10% validation set, 10% test set
- Potential improvement: cross validation.

Other details

- Convert data set variables to **tensors**
- Use **iterator** to save memory and loop much faster
- BertForSequenceClassification
- Choose batch size (authors recommend 16 or 32)
  - We tried 2, 4, 8, 16 & 32.
- Optimizer (authors recommend learning rate = 2e-5 & epsilon = 1e-8)
  - We did not fine tune these hyper parameters.
- Epochs (authors recommend 2 to 4)
  - We chose 4.

Overview of runs

- **Version 1: 15 runs**
  - Several runs with 40K complaints, instead of entire balanced data set (106K)

    → Yielded lower avg. model accuracy than data set with 106K, thus version 2 and 3 runs only contained the entire data set.

Run 1

- 2 batches, 1K complaints, sequence length 512

	precision	recall	f1-score	support
Bank account or service	0.43	0.60	0.50	10
Checking or savings account	0.33	0.14	0.20	7
Mortgage	0.00	0.00	0.00	10
Credit card or prepaid card	0.54	0.70	0.61	10
Credit reporting, credit repair services, or other personal consumer reports	0.82	0.82	0.82	11
Debt collection	0.71	0.83	0.77	6
Money transfer, virtual currency, or money service	0.43	0.50	0.46	6
Money transfers	0.00	0.00	0.00	2
Consumer Loan	1.00	0.92	0.96	12
Payday loan, title loan, or personal loan	0.40	0.57	0.47	7
Prepaid card	0.00	0.00	0.00	1
Student loan	0.75	0.75	0.75	8
Vehicle loan or lease	0.40	0.40	0.40	10
accuracy			0.56	100
macro avg	0.45	0.48	0.46	100
weighted avg	0.53	0.56	0.54	100

Run 2

- 4 batches, 1K complaints, sequence length 128 → Down, because sequence length down

	precision	recall	f1-score	support
Bank account or service	1.00	0.40	0.57	5
Checking or savings account	0.50	0.75	0.60	8
Consumer Loan	0.50	0.08	0.13	13
Credit card or prepaid card	0.73	0.73	0.73	11
Credit reporting, credit repair services, or other personal consumer reports	0.27	0.40	0.32	10
Debt collection	0.67	0.40	0.50	10
Money transfer, virtual currency, or money service	0.40	0.50	0.44	4
Money transfers	0.00	0.00	0.00	2
Mortgage	0.90	0.90	0.90	10
Payday loan, title loan, or personal loan	0.36	0.50	0.42	8
Prepaid card	0.00	0.00	0.00	2
Student loan	0.78	0.88	0.82	8
Vehicle loan or lease	0.29	0.56	0.38	9
accuracy			0.52	100
macro avg	0.49	0.47	0.45	100
weighted avg	0.55	0.52	0.50	100

Went down

Run 3

- 4 batches, 4K complaints, sequence length 128 → Up due to increase in complaints

	precision	recall	f1-score	support
Bank account or service	0.19	0.10	0.13	30
Checking or savings account	0.40	0.50	0.44	34
Consumer Loan	0.47	0.43	0.45	35
Credit card or prepaid card	0.67	0.82	0.73	44
Credit reporting, credit repair services, or other personal consumer reports	0.67	0.81	0.73	32
Debt collection	0.71	0.71	0.71	42
Money transfer, virtual currency, or money service	0.54	0.52	0.53	25
Money transfers	0.00	0.00	0.00	5
Mortgage	0.87	0.89	0.88	37
Payday loan, title loan, or personal loan	0.69	0.71	0.70	35
Prepaid card	1.00	0.10	0.18	10
Student loan	0.80	0.95	0.87	38
Vehicle loan or lease	0.57	0.52	0.54	33
accuracy			0.63	400
macro avg	0.58	0.54	0.53	400
weighted avg	0.61	0.63	0.61	400

Run 4

- 4 batches, 8K complaints, sequence length 128

	precision	recall	f1-score	support
Bank account or service	0.52	0.34	0.41	88
Checking or savings account	0.44	0.48	0.46	75
Consumer Loan	0.37	0.25	0.29	77
Credit card or prepaid card	0.59	0.78	0.67	69
Credit reporting, credit repair services, or other personal consumer reports	0.82	0.72	0.77	80
Debt collection	0.69	0.70	0.70	74
Money transfer, virtual currency, or money service	0.65	0.68	0.67	69
Prepaid card	0.00	0.00	0.00	11
Mortgage	0.81	0.89	0.85	74
Payday loan, title loan, or personal loan	0.51	0.63	0.56	49
Money transfers	0.38	0.50	0.43	6
Student loan	0.84	0.89	0.86	63
Vehicle loan or lease	0.56	0.63	0.59	65
accuracy			0.62	800
macro avg	0.55	0.58	0.56	800
weighted avg	0.61	0.62	0.61	800

Run 5

- 4 batches, 20K complaints, sequence length 128 → Up due to increase in complaints

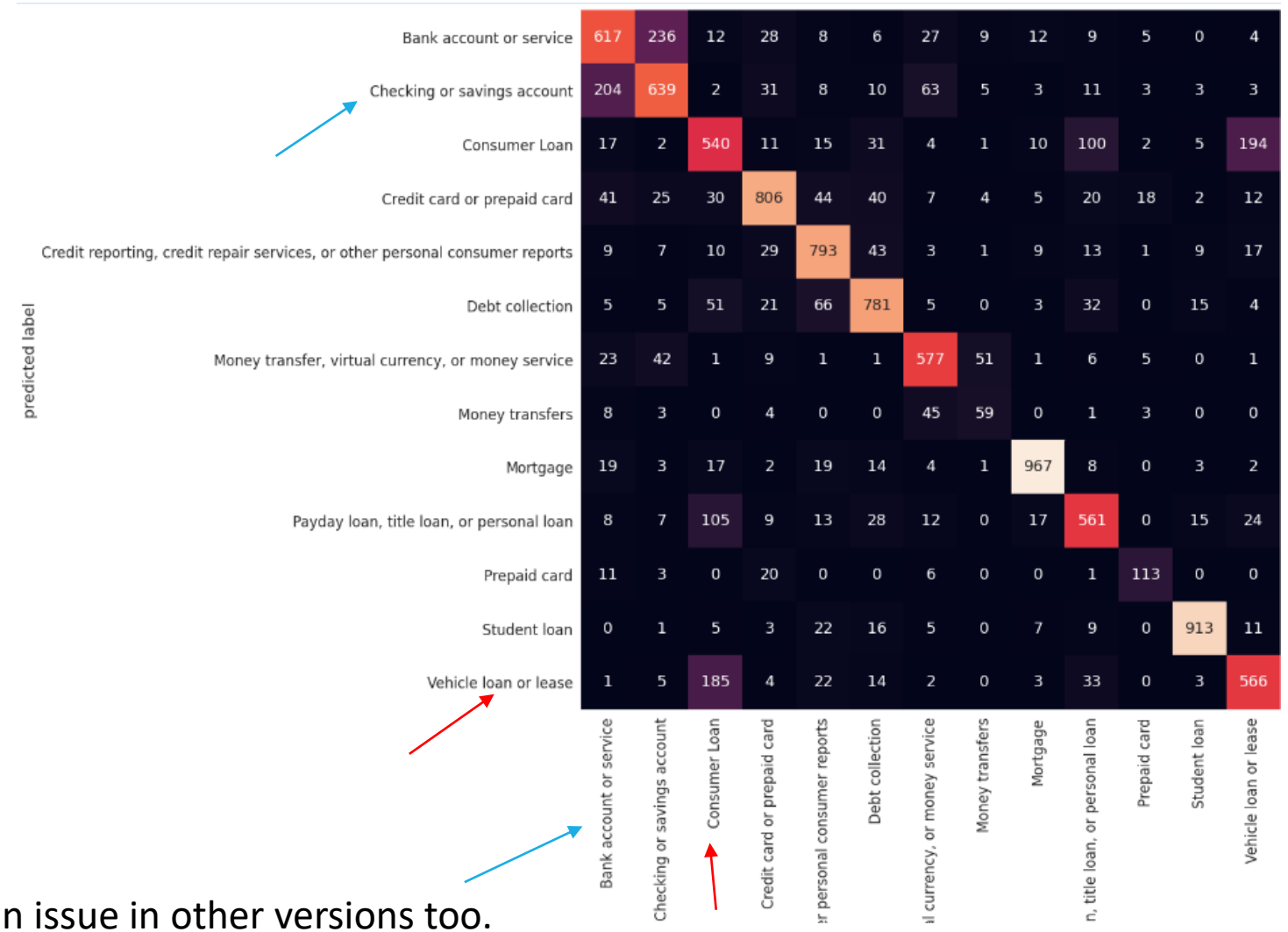
	precision	recall	f1-score	support
Bank account or service	0.54	0.53	0.53	190
Checking or savings account	0.55	0.55	0.55	193
Consumer Loan	0.48	0.46	0.47	192
Credit card or prepaid card	0.74	0.81	0.77	214
Credit reporting, credit repair services, or other personal consumer reports	0.78	0.75	0.77	181
Debt collection	0.71	0.73	0.72	190
Money transfer, virtual currency, or money service	0.67	0.63	0.65	139
Money transfers	0.14	0.07	0.09	29
Mortgage	0.84	0.89	0.87	168
Payday loan, title loan, or personal loan	0.66	0.66	0.66	159
Prepaid card	0.72	0.91	0.81	23
Student loan	0.91	0.90	0.90	183
Vehicle loan or lease	0.60	0.63	0.62	139
accuracy			0.68	2000
macro avg	0.64	0.66	0.65	2000
weighted avg	0.67	0.68	0.68	2000

Run 8

- 4 batches, 106K (all) complaints, sequence length 256

	precision	recall	f1-score	support
Bank account or service	0.63	0.64	0.64	963
Checking or savings account	0.65	0.65	0.65	978
Consumer Loan	0.58	0.56	0.57	958
Credit card or prepaid card	0.76	0.82	0.79	977
Credit reporting, credit repair services, or other personal consumer reports	0.84	0.78	0.81	1011
Debt collection	0.79	0.79	0.79	984
Money transfer, virtual currency, or money service	0.80	0.76	0.78	760
Money transfers	0.48	0.45	0.46	131
Mortgage	0.91	0.93	0.92	1037
Payday loan, title loan, or personal loan	0.70	0.70	0.70	804
Prepaid card	0.73	0.75	0.74	150
Student loan	0.92	0.94	0.93	968
Vehicle loan or lease	0.68	0.68	0.68	838
accuracy			0.75	10559
macro avg	0.73	0.73	0.73	10559
weighted avg	0.75	0.75	0.75	10559

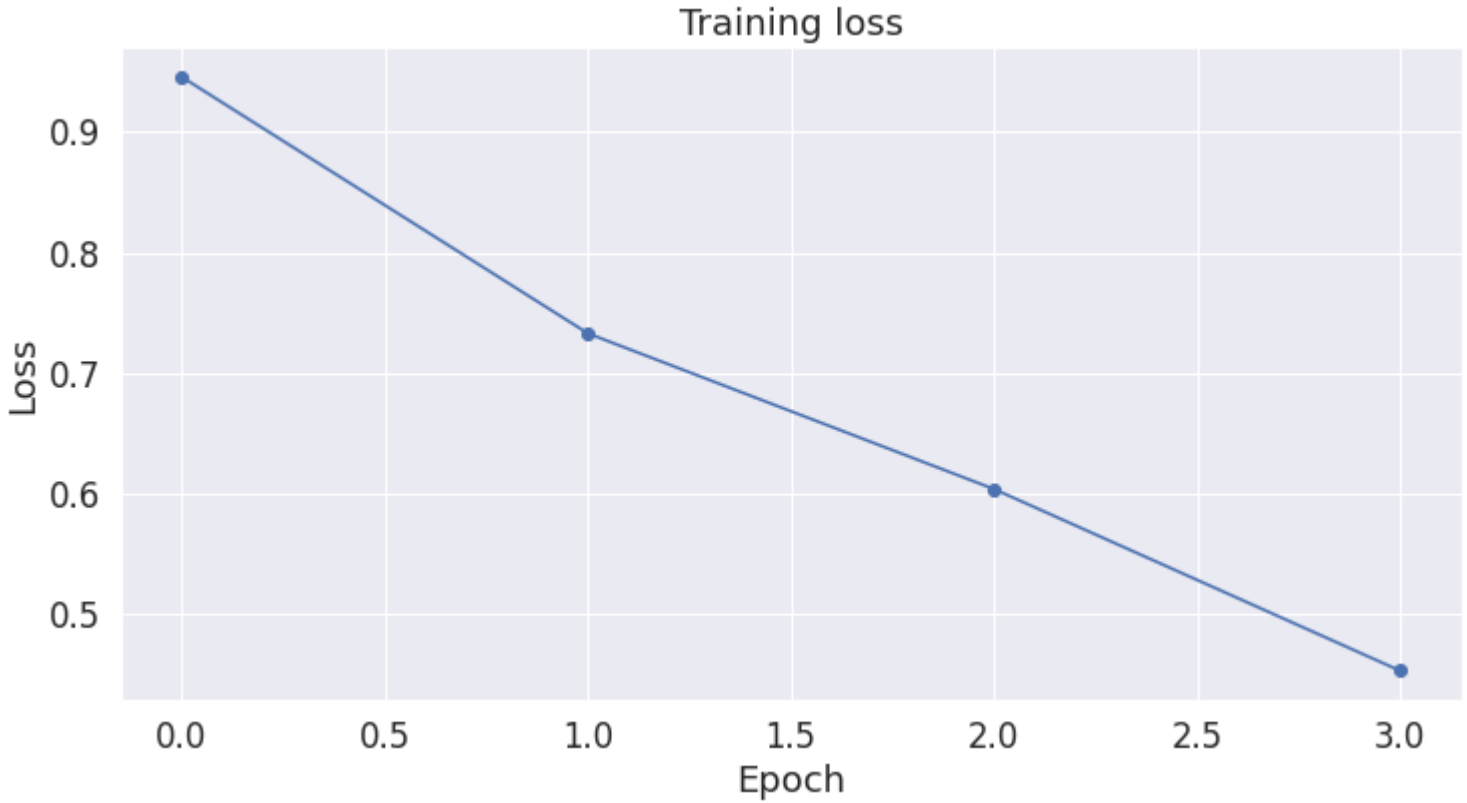
Run 8: confusion matrix



Confusion remains an issue in other versions too.

Run 8: training loss

- Training loss below 0.5 for first time.



Runs 9-15

- **run 9)** 8 batches, 40K complaints, sequence length 128: avg. accuracy 0.71
- **run 10)** 8 batches, 106K (all) complaints, sequence length 128: avg. accuracy 0.74
- **run 11)** 8 batches, 40K complaints, sequence length 256: avg. accuracy 0.74
- **run 12)** 8 batches, 106K (all) complaints, sequence length 256: avg. accuracy 0.75
- **run 13)** 16 batches, 106K (all) complaints, sequence length 128: avg. accuracy 0.74
- **run 14)** 16 batches, 40K complaints, sequence length 256: avg. accuracy 0.72
- **run 15)** 16 batches, 106K (all) complaints, sequence length 256: avg. accuracy 0.75

Observations:

More complaints yield slightly higher results.

Increase in batch size =/= better performance, despite intuition it might.

Version 2 runs (nearly fully cleaned)

- **run 1)** 8 batches, 106K (all) complaints, sequence length 128: avg. accuracy 0.75
- **run 2)** 8 batches, 106K (all) complaints, sequence length 256: avg. accuracy 0.76
- **run 3)** 16 batches, 106K (all) complaints, sequence length 128: avg. accuracy 0.75
- **run 4)** 16 batches, 106K (all) complaints, sequence length 256: avg. accuracy 0.77

Observations:

Steady accuracy, though training loss between epochs is lower than version 1.

Models take less time to run.

Increase in batch size == slightly better result, but negligible.

Version 3 runs (partially cleaned)

- **run 1)** 8 batches, 106K (all) complaints, sequence length 128: avg. accuracy 0.77
- **run 2)** 8 batches, 106K (all) complaints, sequence length 256: avg. accuracy 0.78
- **run 3)** 16 batches, 106K (all) complaints, sequence length 128: avg. accuracy 0.77
- **run 4)** 16 batches, 106K (all) complaints, sequence length 256: avg. accuracy 0.79
- **run 5) 32 batches, 106K (all) complaints, sequence length 256: avg. accuracy 0.78**

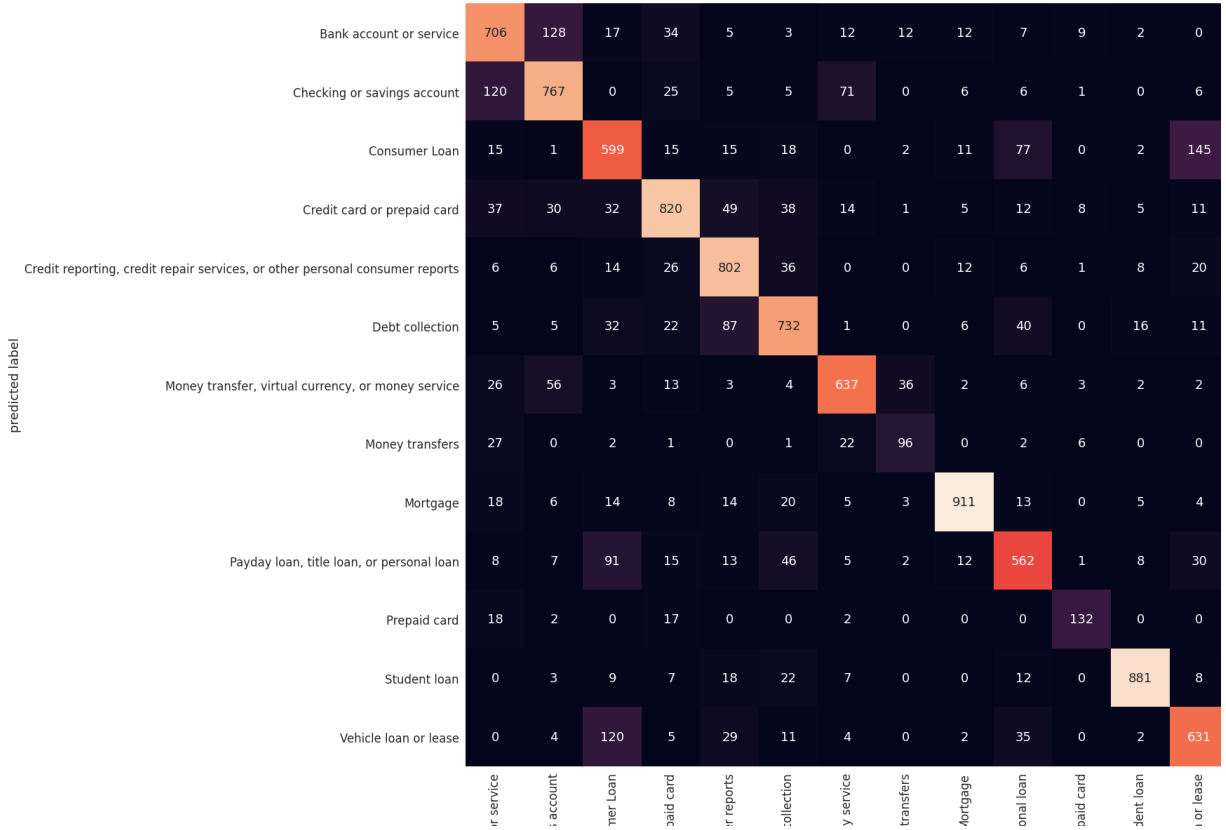
→ Run 5 substantially faster than Run 4

Observations:

Better results than version 1 and 2.
Models take less time to run.
It appears removing words (tokens) is bad for BERT's understanding of context.
Training loss between epochs even lower than version 2.

Version 3, run 5

- 32 batches, 106K (all) complaints, sequence length 256



Version 3, run 5

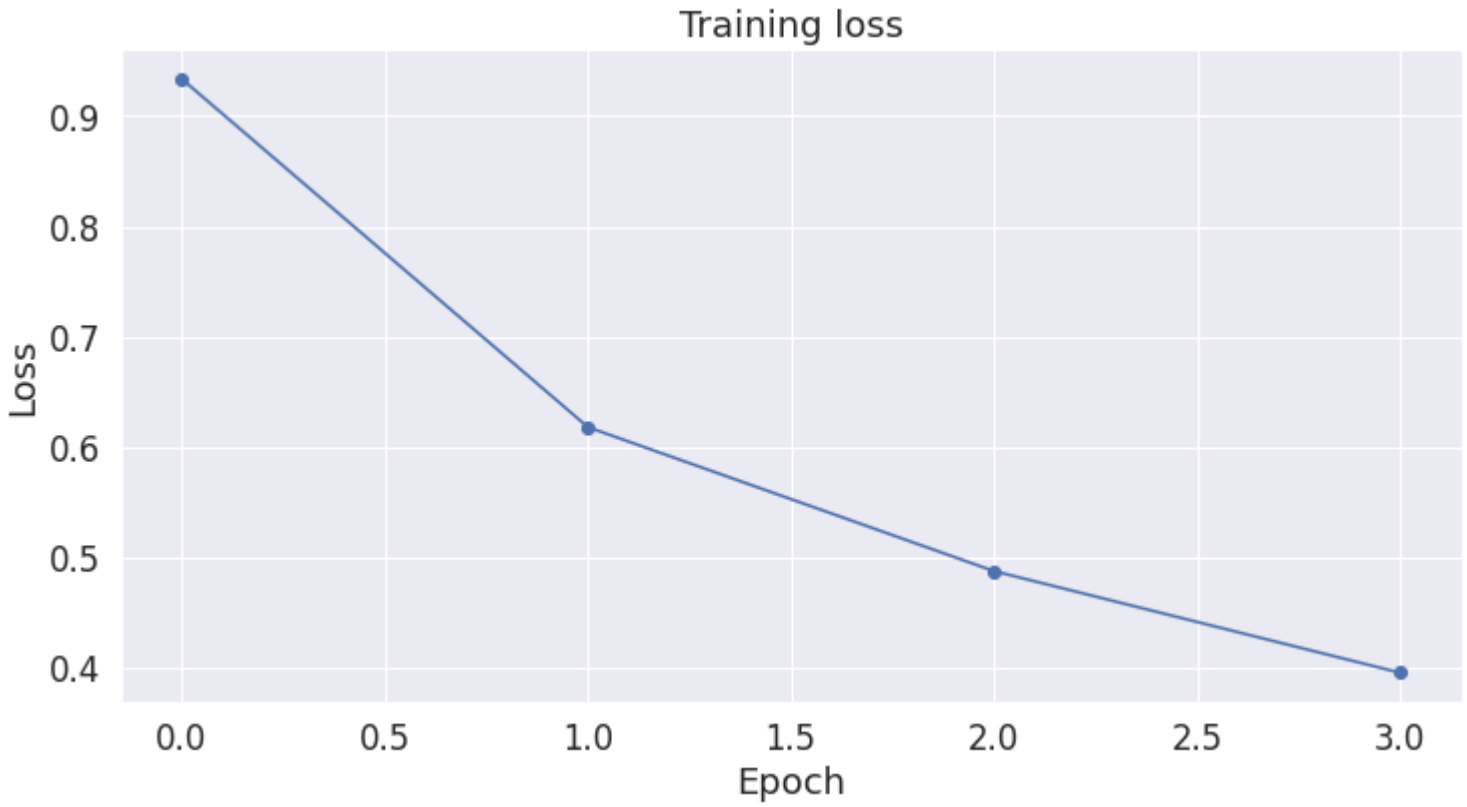
- 32 batches, 106K (all) complaints, sequence length 256

	precision	recall	f1-score	support
Bank account or service	0.75	0.72	0.73	986
Checking or savings account	0.76	0.76	0.76	1015
Consumer Loan	0.67	0.64	0.65	933
Credit card or prepaid card	0.77	0.81	0.79	1008
Credit reporting, credit repair services, or other personal consumer reports	0.86	0.77	0.81	1040
Debt collection	0.76	0.78	0.77	936
Money transfer, virtual currency, or money service	0.80	0.82	0.81	780
Money transfers	0.61	0.63	0.62	152
Mortgage	0.89	0.93	0.91	979
Payday loan, title loan, or personal loan	0.70	0.72	0.71	778
Prepaid card	0.77	0.82	0.80	161
Student loan	0.91	0.95	0.93	931
Vehicle loan or lease	0.75	0.73	0.74	868
accuracy			0.78	10567
macro avg	0.77	0.77	0.77	10567
weighted avg	0.78	0.78	0.78	10567

Consolidate even more products?
Not if distinction is required between credit & prepaid cards.

Version 3, run 5

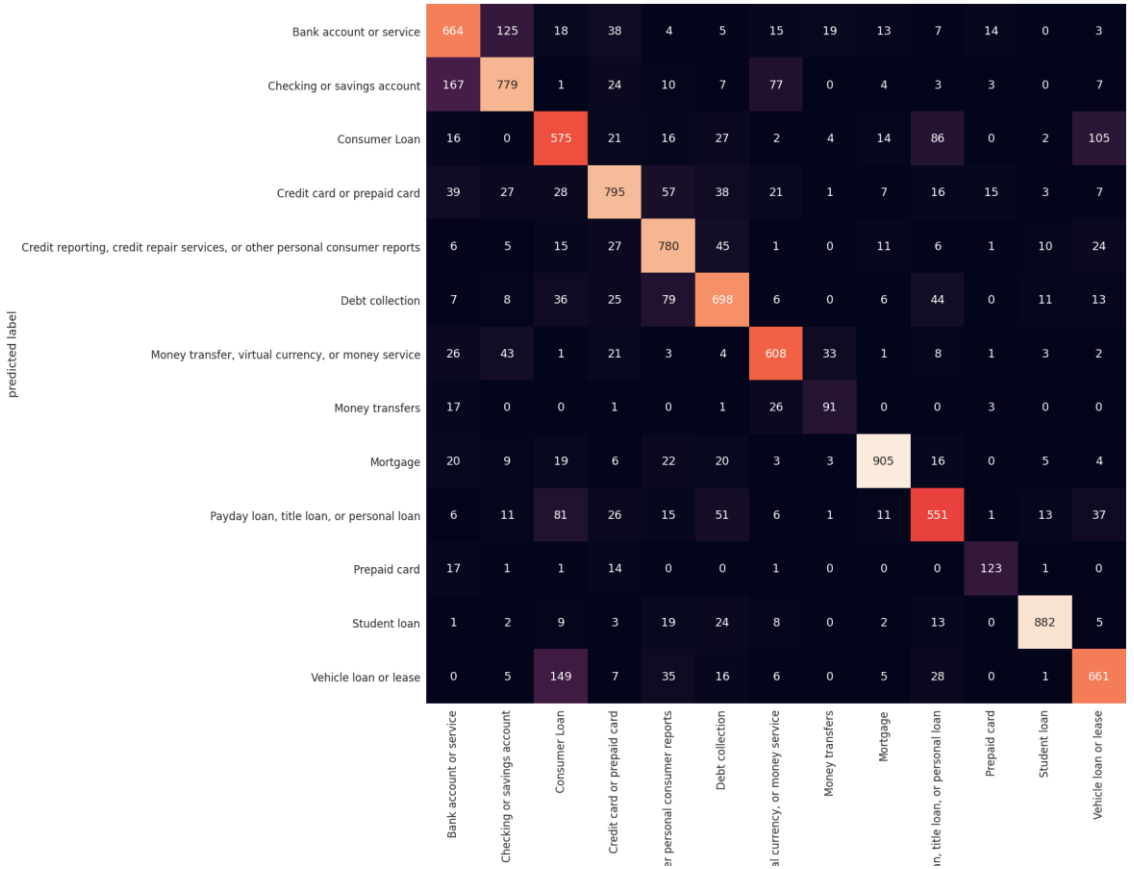
- 32 batches, 106K (all) complaints, sequence length 256



→ More epochs?

Version 3, run 6

- 32 batches, 106K (all) complaints, sequence length 256, **epochs = 7**



Same issues as before.

Version 3, run 6

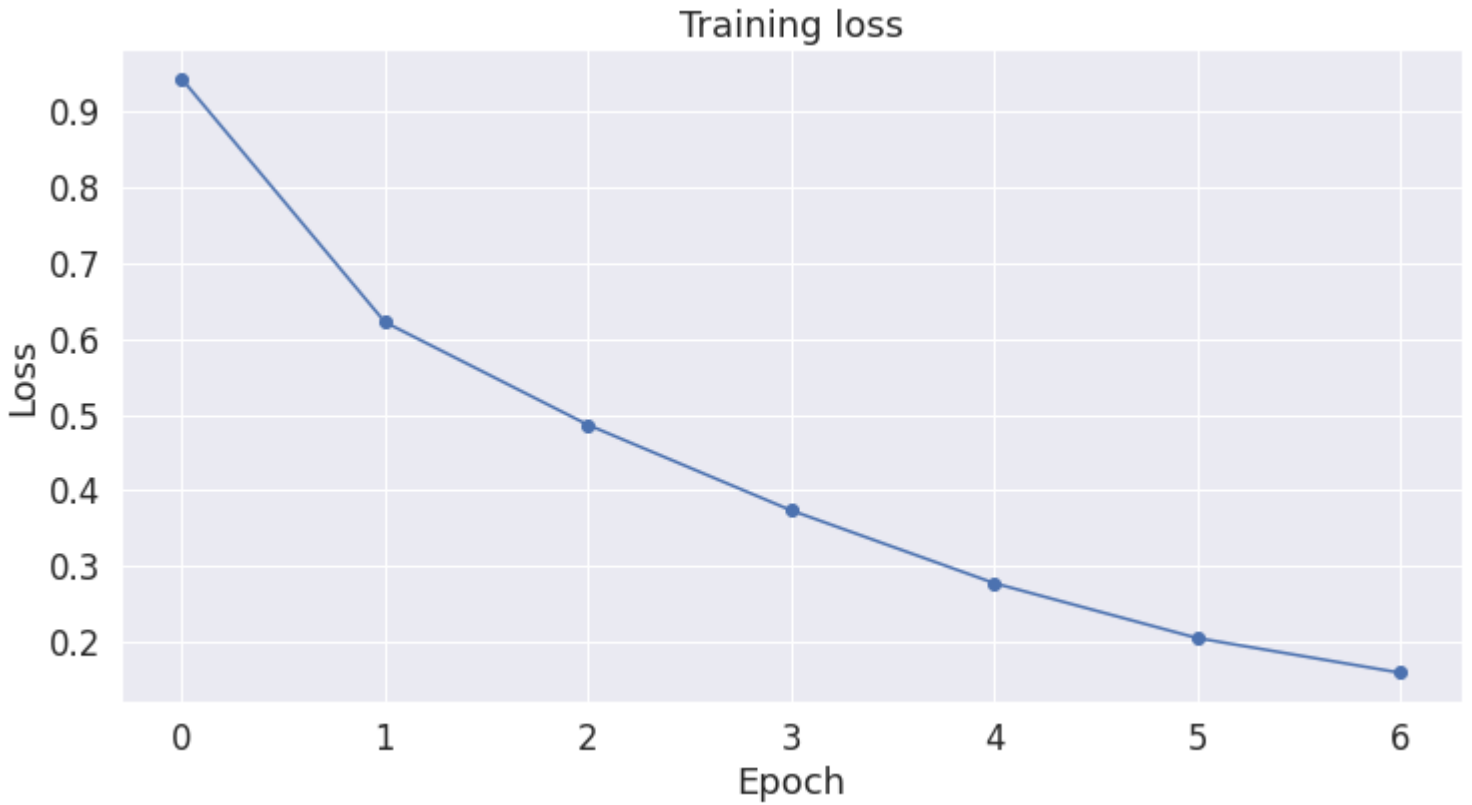
- 32 batches, 106K (all) complaints, sequence length 256, epochs = 7

	precision	recall	f1-score	support
Bank account or service	0.72	0.67	0.69	986
Checking or savings account	0.72	0.77	0.74	1015
Consumer Loan	0.66	0.62	0.64	933
Credit card or prepaid card	0.75	0.79	0.77	1008
Credit reporting, credit repair services, or other personal consumer reports	0.84	0.75	0.79	1040
Debt collection	0.75	0.75	0.75	936
Money transfer, virtual currency, or money service	0.81	0.78	0.79	780
Money transfers	0.65	0.60	0.63	152
Mortgage	0.88	0.92	0.90	979
Payday loan, title loan, or personal loan	0.68	0.71	0.69	778
Prepaid card	0.78	0.76	0.77	161
Student loan	0.91	0.95	0.93	931
Vehicle loan or lease	0.72	0.76	0.74	868
accuracy			0.77	10567
macro avg	0.76	0.76	0.76	10567
weighted avg	0.77	0.77	0.77	10567

More epochs =/= model fine-tuning improvements

Version 3, run 6

- 32 batches, 106K (all) complaints, sequence length 256, epochs = 7



Conclusion

- Batch size does not necessarily improve model, but does make fine-tuning faster.
- Sequence length of complaint (128 → 256): improves avg. model accuracy only slightly

  → BERT does not need as many words to understand complaint.

- Hypothesis was: fully cleaned version (1) > version 2 & 3

  → Proven wrong! Version 3 (with censored data (XXXX)) is superior. Potentially due to sentence structure.

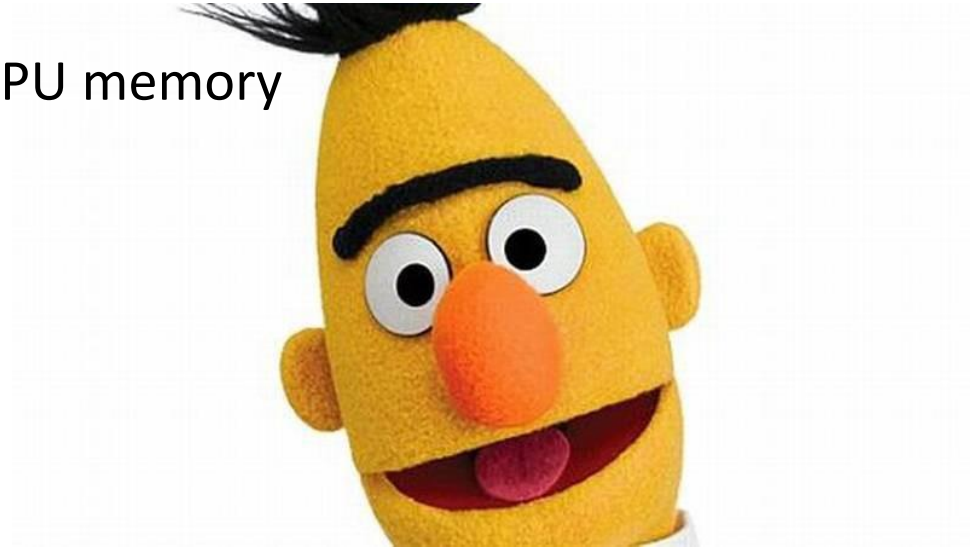
- Could potentially push to 80+ % avg. model accuracy, if loans are grouped as one product, or a better distinction can be made/added (more information for BERT).

Conclusion

- **BERT** is indeed <u>much better</u> when compared to RF (0.69) or SVM (0.58) in terms of avg. model accuracy.
  → ***Narrative*** only

- Additional experiment: more epochs. Improvement? → Not really.

- Large BERT → needs more GPU memory



Motivated to keep 'learning'!