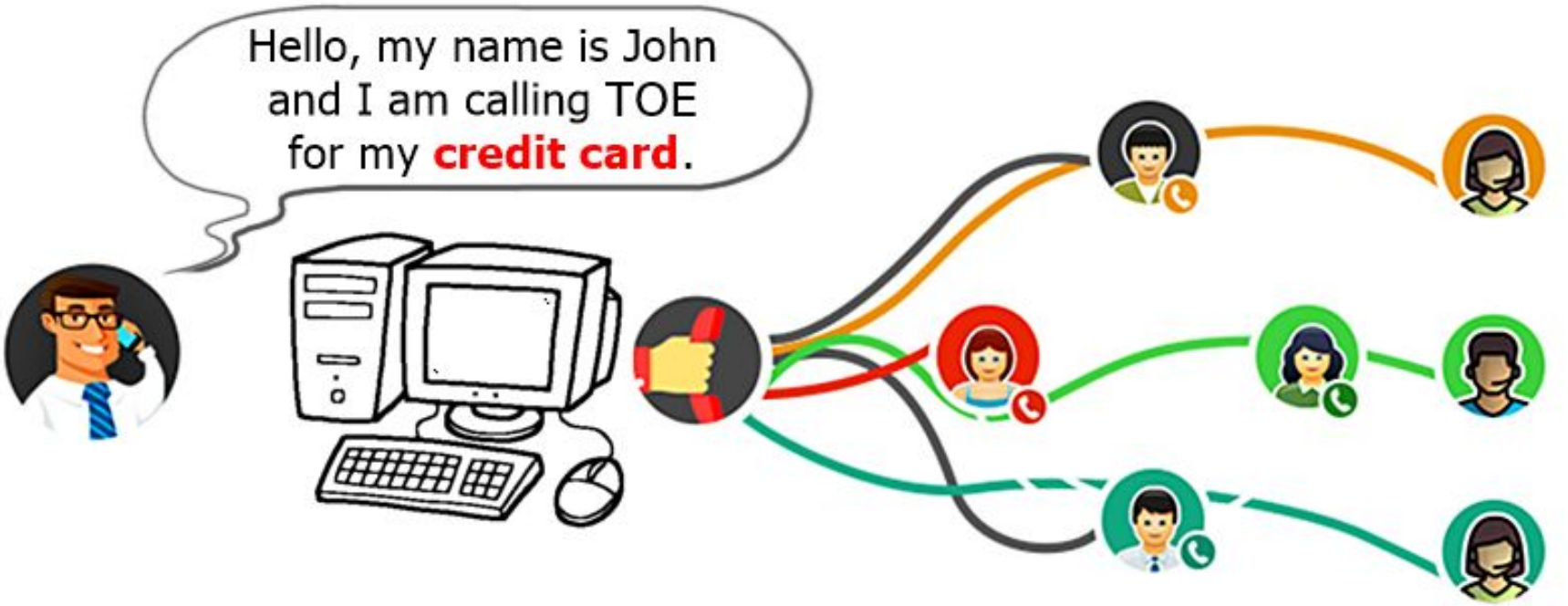
15.2 Call Routing System

Πολυτεχνείο Κρήτης
Σχολή Ηλεκτρολόγων Μηχανικών & Μηχανικών Η/Υ
Στατιστική Μοντελοποίηση και Αναγνώριση Προτύπων (ΤΗΛ 311)
Εαρινό Εξάμηνο 2018

Μαρία Αποστολίδου, 2012030114
Νικόλας Κυπαρισσάς, 2012030112

Problem: Call Routing



Solution

Text classification using Gaussian Mixture Models (GMMs)

Form of Data

Class	Text
2	ah credit card six seven nine one five ...
8	card two oh four eight two two ask on ...
2	like to make a credit card call
3	hi I'd like to make a call to london and ...
2	hi I wanna use my credit card
5	yes I'd like to place a long distance ...
2	yes I'm trying to call a long distance ...
9	yes operator I'd like to make a time I ...
...	...

Data Representation

Example data:

```
I’d like some cake
I’d like to buy some cake
```

Dictionary:

I’d	like	some	cake	to	buy	a

Term Document:

d1	1	1	1	1	0	0	0
d2	1	1	0	1	1	1	1

Data Representation

$$TF(t) = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document}$$

Dictionary:

I'd	like	some	cake	to	buy	a

Term Frequency (TF):

d1	1/4	1/4	1/4	1/4	0	0	0
d2	1/6	1/6	0	1/6	1/6	1/6	1/6

Data Representation

$$IDF(t) = ln\left(\frac{Total\ number\ of\ documents}{Total\ number\ of\ documents\ with\ term\ t\ in\ it}\right)$$

Dictionary:

I'd	like	some	cake	to	buy	a

Inverse Document Frequency (IDF):

ln(1) =0	ln(1) =0	ln(2) =0.69	ln(1) =0	ln(2) =0.69	ln(2) =0.69	ln(2) =0.69

Data Representation

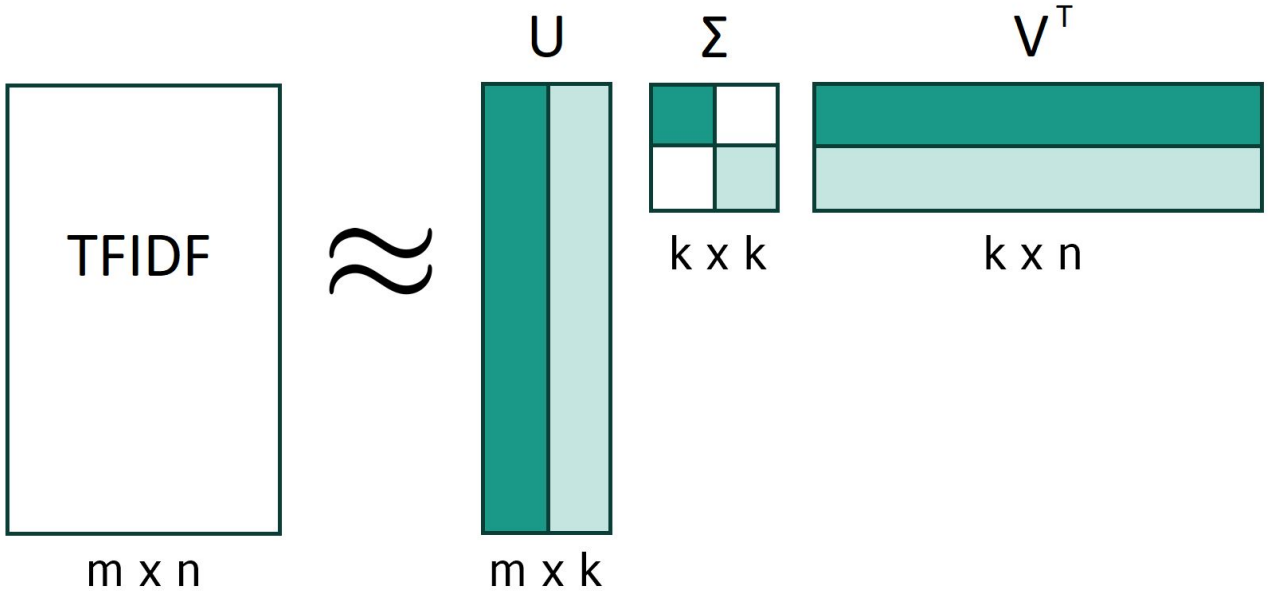
Dict:	I'd	like	some	cake	to	buy	a
TF:							
d1	1/4	1/4	1/4	1/4	0	0	0
d2	1/6	1/6	0	1/6	1/6	1/6	1/6
IDF:	0	0	0.69	0	0.69	0.69	0.69
TFIDF:	0	0	0.1725	0	0	0	0
	0	0	0	0	0.115	0.115	0.115

Feature Dimensionality Reduction

- Sort *dictionary*, *TF* and *IDF* according to *TFIDF* values
- Discard the least important terms from *dictionary*, *TF* and *IDF*

Feature Extraction

- Dimensionality Reduction of *TFIDF* with Truncated Singular-value Decomposition (SVD)



Gaussian Mixture Models (GMMs)

- Gather the corresponding TFIDF samples for each class

- A GMM is trained for each class to estimate a distribution over the features

Classification

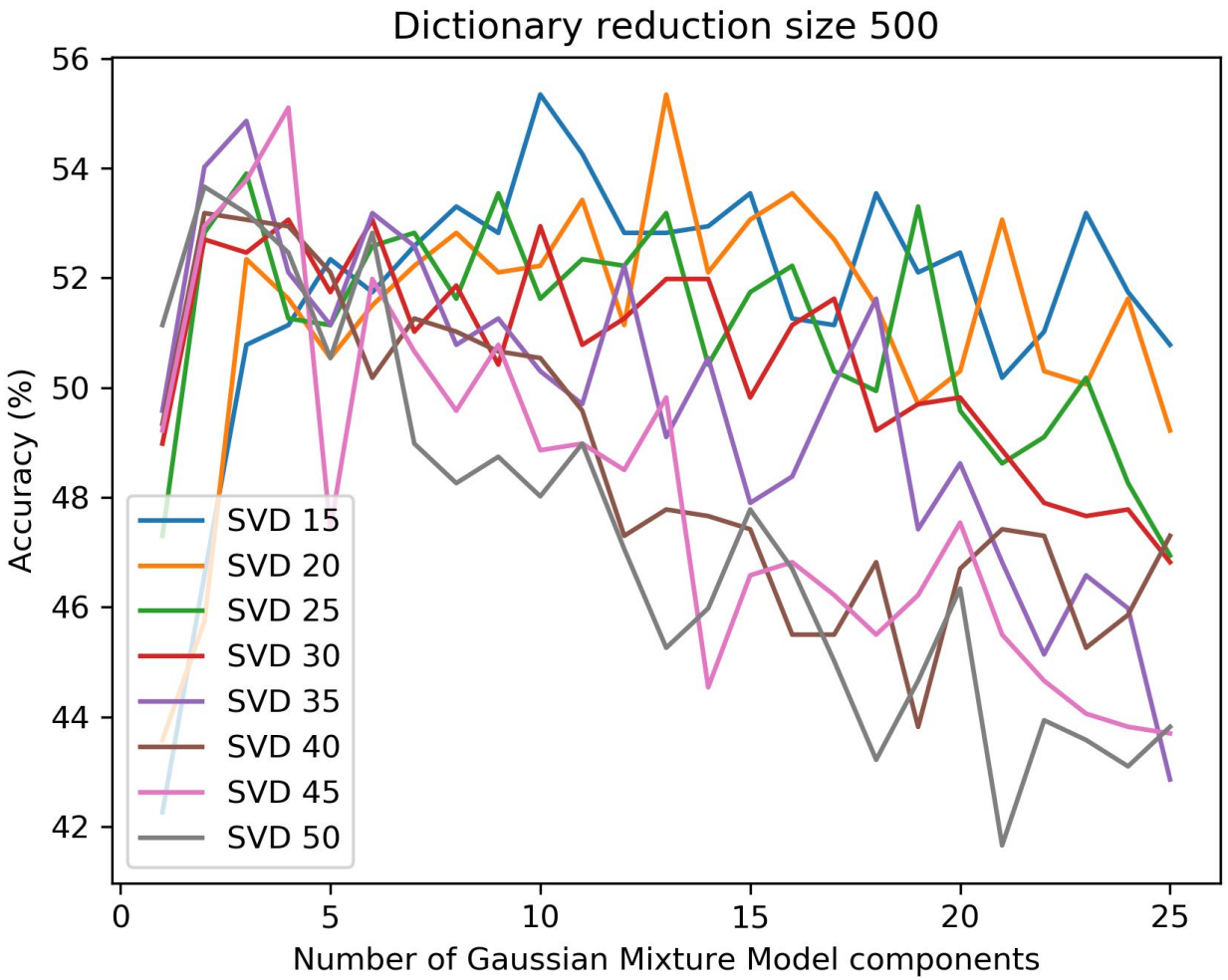
- Transform the test samples to the dimensions of the feature space
- Calculate the loglikelihood probabilities for each GMM
- Find the maximum loglikelihood probability
- Cassify test samples to the classes for which the loglikelihood probability is maximum

Testing

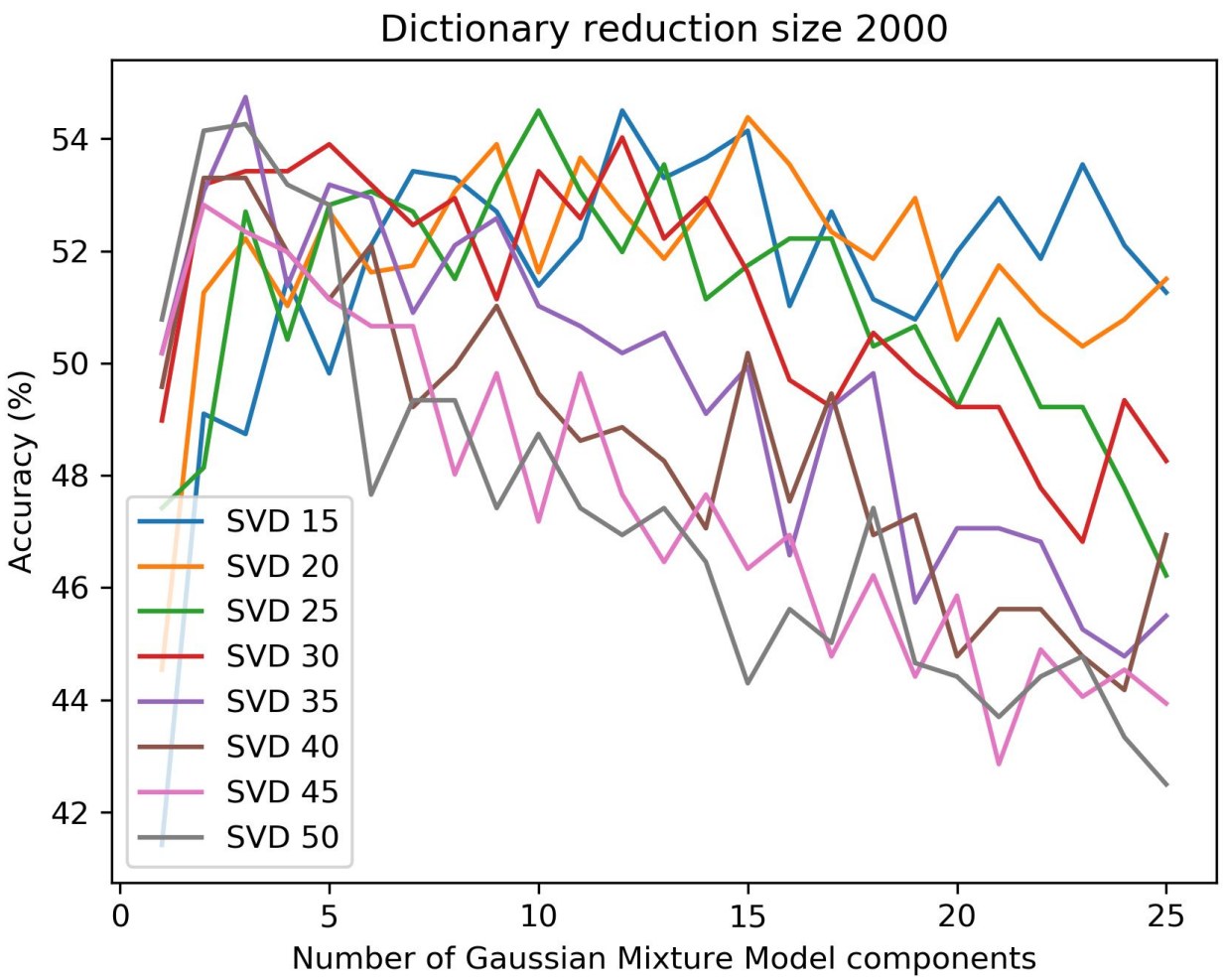
We examined different values for the parameters of:

- initial word reduction
- number of SVD components
- number of each GMM's components

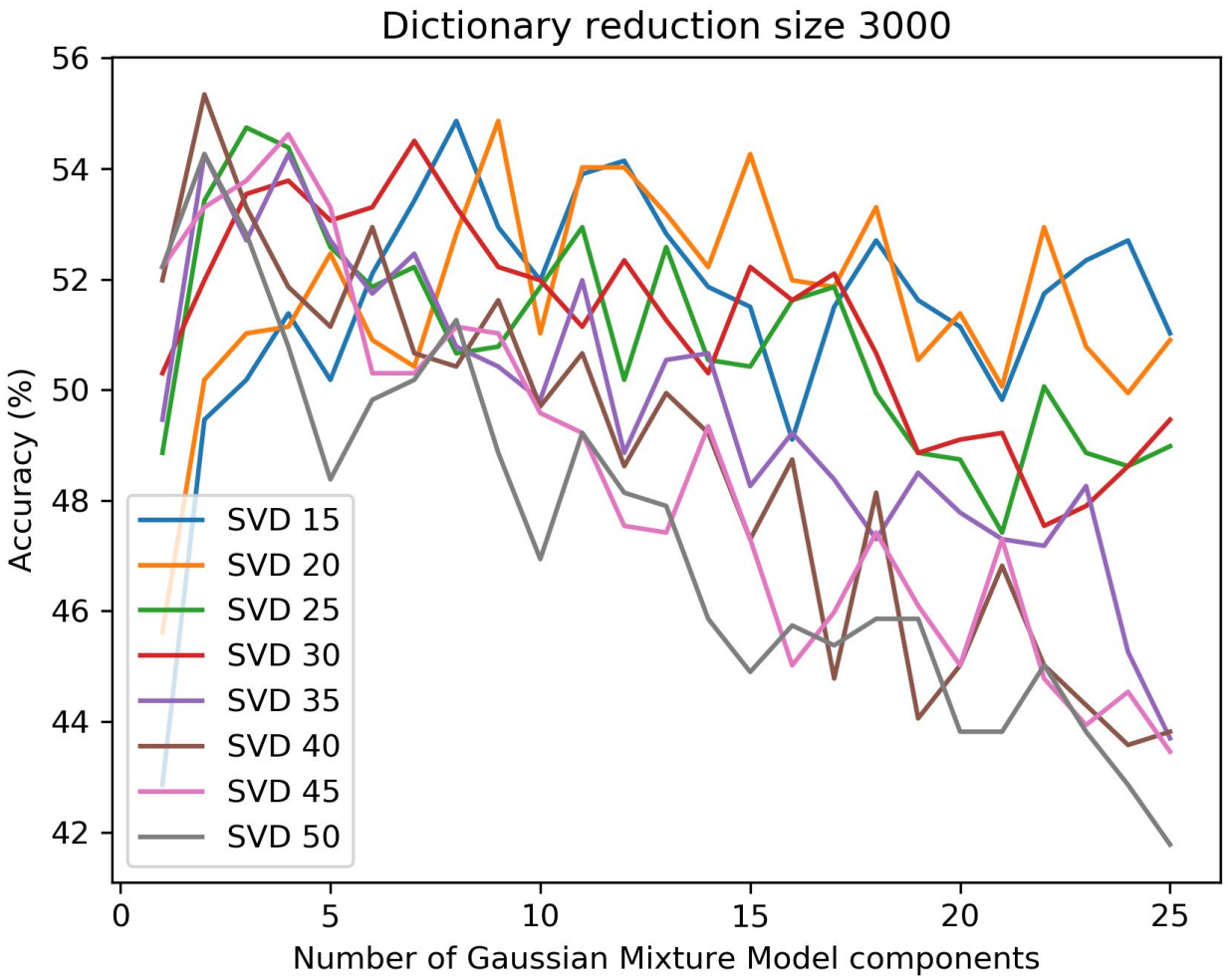
Comparative Results



Comparative Results



Comparative Results



State of the art - Future work

- Naive Bayes
- SVM, cosine similarity
- Edit Distance :
    i. Find a representative super sentence for each class
    ii. classify samples to the class for which the edit distance between respresentative and sample sentence is minimum

Thank you for your attention!