Text Mining Applied to Clinical Notes

Ayoub Bagheri

Utrecht University, UMC Utrecht

a.bagheri@uu.nl; a.bagheri-2@umcutrecht.nl

Kick-off meeting Special Interest Group Text Mining,

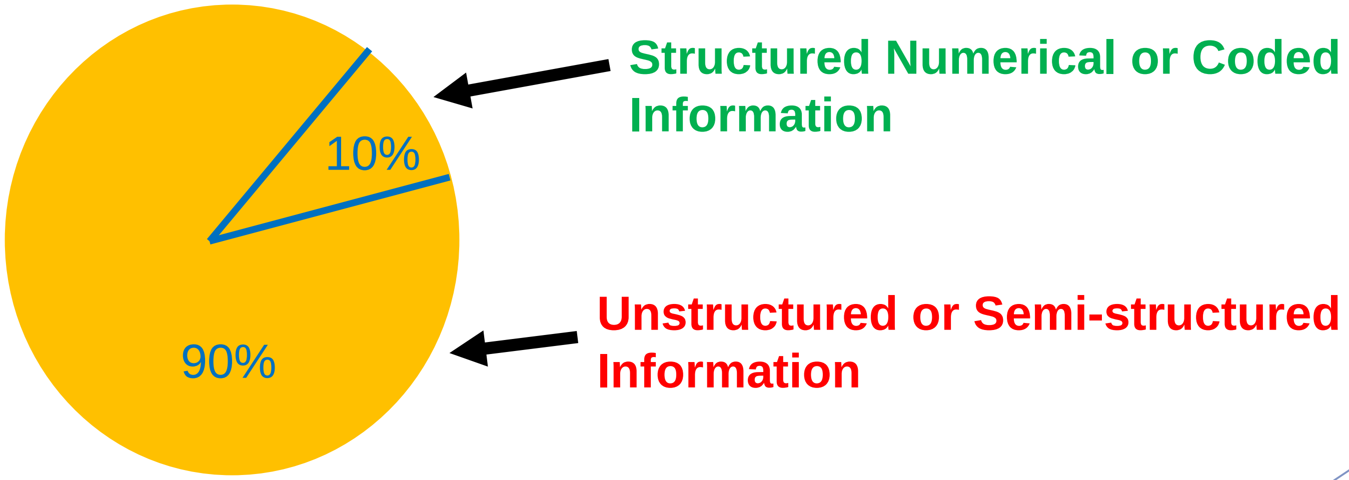
Utrecht University, Utrecht, September 2018.

Outline

- Text mining
- Applications
- Challenges
- Unsupervised enrichment framework
- Topic models
- ETM, Results

Text mining

- Text mining is the art of turning free text or speech into numerical variables and then mining them with statistical techniques and learning algorithms.
- **It is impossible to read all of the information, text mining does this for us!**



Structured Numerical or Coded Information

Unstructured or Semi-structured Information

Text mining

- Text mining is about extracting the useful information, patterns and knowledge from text for use in decision support and estimation.
- Text Mining is a field at the intersection of
  - Computer science
  - Artificial intelligence
  - And linguistics

The ipod is so small! ☺
The monitor is so small! ☹

I made her duck.

1. I cooked *duck* for her.
2. I cooked *duck* belonging to her.
3. I created a toy duck which she owns.
4. I caused her to quickly lower her head or body.
5. I used magic and turned her into a *duck*.

Applications

- Spell checking, keyword search, finding synonyms
- Extracting information from websites such as Product prices
- Social media analysis
- Sentiment analysis
- Spam filtering
- Machine translation
- Question answering

Mining healthcare data

Diagnosis:

Recognize and classify patterns in patient profiles and attributes

Therapy:

Select from available treatment methods; based on effectiveness, suitability to patient, etc.

Prognosis:

Predict future outcomes based on previous experience and present conditions (history of patient, etc.)

Healthcare applications

- Monitoring
- Biomedical/Biological Analysis
- Epidemiological Studies
- Hospital Management
- Medical Instruction and Training

Clinical text datasets

- MIMIC-III
  - Openly available dataset developed by the MIT Lab for Computational Physiology
  - Deidentified health data associated with >40,000 critical care patients
  - In addition to structured clinical data (demographics, vital signs, laboratory tests, medications, etc.), it contains over 2 million free-text notes from nurses, physicians, specialists, and more
- i2b2 NLP Research Data Sets
  - The i2b2 (Informatics for Integrating Biology and the Bedside) project has organized a yearly series of shared tasks, starting in 2006
  - Deidentified notes from the Research Patient Data Repository at Partners HealthCare
  - De-identification, named entity and relation extraction, negation and modality, co-reference resolution, temporal information extraction

Clinical text datasets

- ShARe/CLEF eHealth
  - The Sharing Annotated Resources (ShARe) / Conference and Labs of the Evaluation Forum (CLEF)
  - Shared tasks on disease/disorder named entity recognition, normalization of named entities
  - Unified Medical Language System (UMLS), and disease/disorder template filling (2013-2016)
- SemEval
  - Several shared tasks in the clinical domain have been organized as a part of the yearly SemEval competitions.
- MedNLPDoc
  - Medical Natural Language Processing for Clinical Document (MedNLPDoc)
  - Processing of Japanese clinical records
  - Named entity recognition, term normalization, and International Codes for Diseases (ICD) disease name identification

Challenges in medical text mining

- Short text
  - Sparsity
  - Dependency to external information
- High level of noise
  - Large number of unknown words, non-words and poor grammatical sentences
  - Complex medical vocabularies
  - Misspellings
  - Acronyms and abbreviations
  - Unknown nonwords are generally the clinical patterns including scores and measures
- Imbalanced data distribution
- Negations, reference resolution
- Lack of dictionaries
- Validation

An unsupervised enrichment framework for classifying clinical cardiovascular notes

Ayoub Bagheri, Daniel Oberski, Peter G.M. van der Heijden, Arjan Sammani, and Folkert W. Asselbergs

***Abstract*—Given the rate at which text data is digitally gathered in the medical field, there is a growing need for automated tools which can analyze and classify free-text in the medical domain. One supreme challenge in automatically analyzing clinical notes is the sparsity of short texts and dependency to external information. In this paper, to tackle the data sparsity issue we propose a novel Natural Language Processing (NLP) toolkit, based on an unsupervised algorithm, the latent Dirichlet allocation method, to obtain semantic representations of short texts by leveraging topic clusters information and internal knowledge acquisition. Using Dutch clinical cardiovascular notes along with experiments on unlabeled text data of the EHR from the UMCU hospital in Utrecht, the Netherlands, the results show that: (i) the enriched data representation significantly**

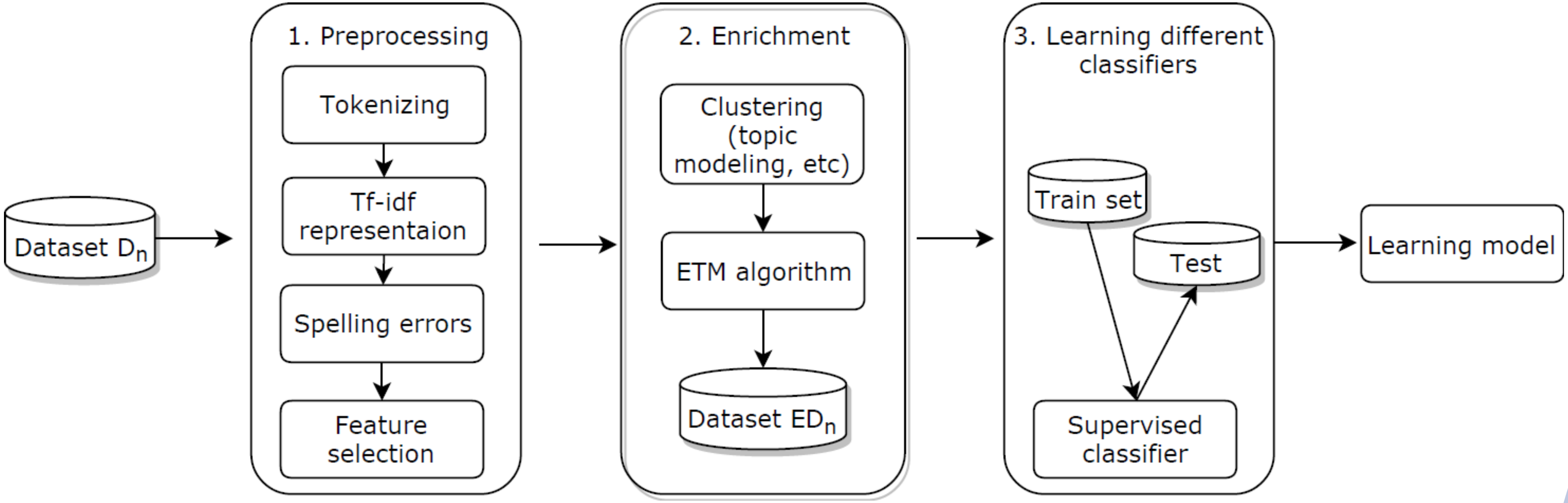
Hence, there is a need for automatic NLP tools to mine such texts with the aim to extract implicit, previously unknown, and potentially useful information from data, with a plausible accuracy on results. Many researchers have been addressing the task of mining clinical text data for different applications in the healthcare area [1], [2], [4], [6], [7], [8], approaching it as a basic text classification problem.

Two major challenges in clinical text mining are the unstructured characteristic of free-textual datasets, and the sparsity presented in short medical text. Short texts refer to texts with limited context, where the sparsity of content makes text mining difficult [9], [10], [11], [12]. The very small counts

Enrichment model for classifying clinical notes

- Semantic representations of short texts
  - Need for organizing and classifying text
- Leveraging topic clusters
  - Sparsity of short text data
  - Latent Dirichlet Allocation
- Enriching short text representation, internal knowledge acquisition
  - Using mixture of the hidden topics
  - Lack external knowledge repositories in medical domain
  - Lack of resources in different languages (Dutch)

Proposed framework

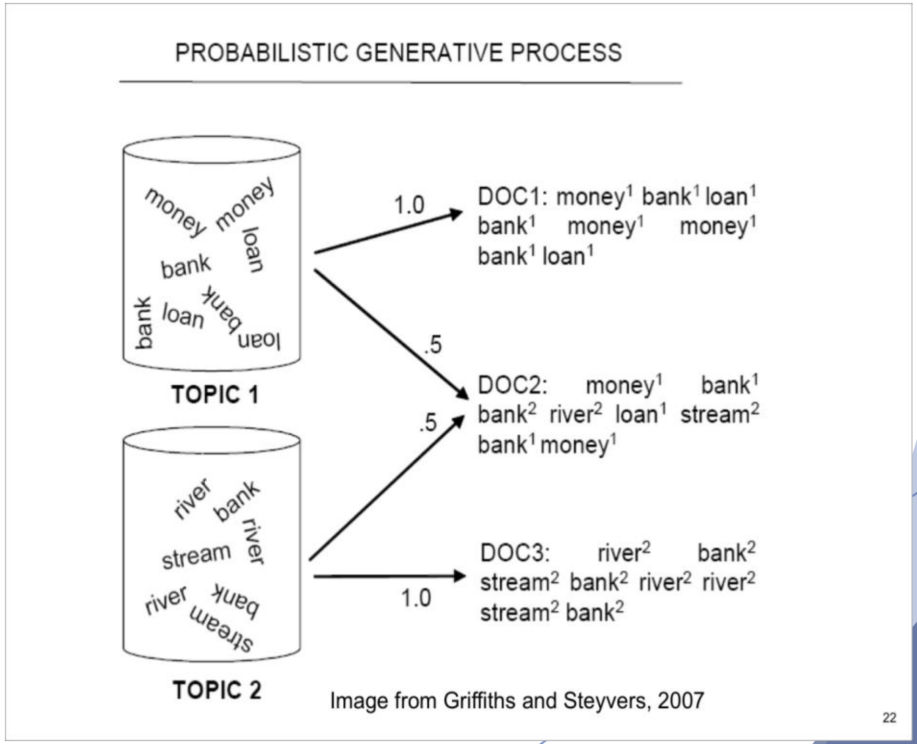


Topic models

- Three concepts: words, topics, and documents
- Documents are a collection of words and have a probability distribution over topics
- Topics have a probability distribution over words

- Model:
  - Topics made up of words

    used to generate documents



PROBABILISTIC GENERATIVE PROCESS

DOC1: money^1 bank^1 loan^1 bank^1 money^1 money^1 bank^1 loan^1

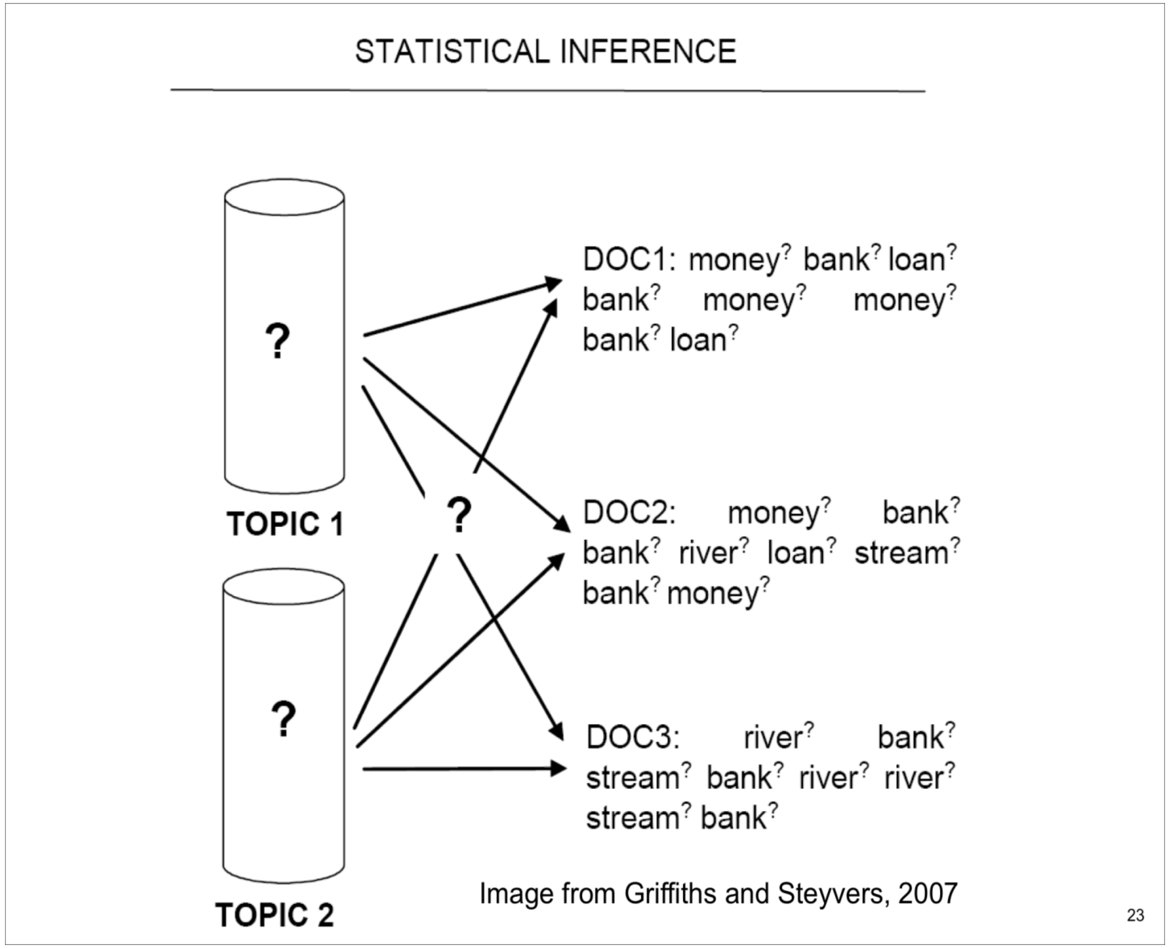
DOC2: money^1 bank^1 bank^2 river^2 loan^1 stream^2 bank^1 money^1

DOC3: river^2 bank^2 stream^2 bank^2 river^2 river^2 stream^2 bank^2

Image from Griffiths and Steyvers, 2007

Topic models (lda or gensim)

Reality: Documents observed, infer topics



STATISTICAL INFERENCE

DOC1: money$^{?}$ bank$^{?}$ loan$^{?}$ bank$^{?}$ money$^{?}$ money$^{?}$ bank$^{?}$ loan$^{?}$

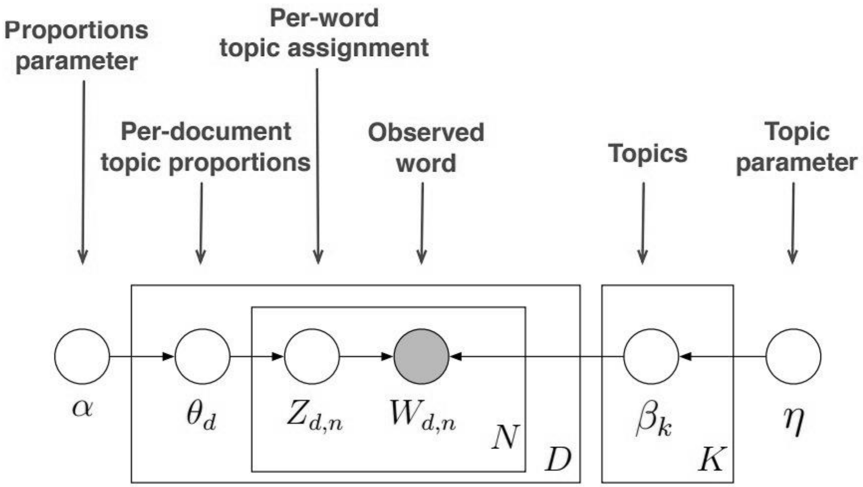
DOC2: money$^{?}$ bank$^{?}$ bank$^{?}$ river$^{?}$ loan$^{?}$ stream$^{?}$ bank$^{?}$ money$^{?}$

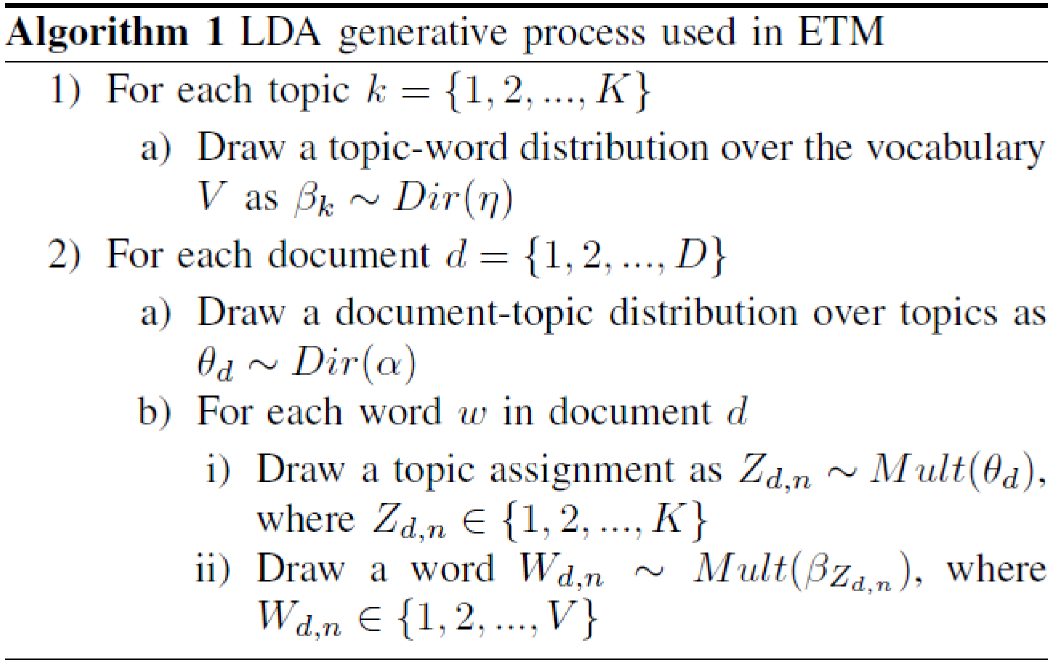
DOC3: river$^{?}$ bank$^{?}$ stream$^{?}$ bank$^{?}$ river$^{?}$ river$^{?}$ stream$^{?}$ bank$^{?}$

Image from Griffiths and Steyvers, 2007

23

LDA

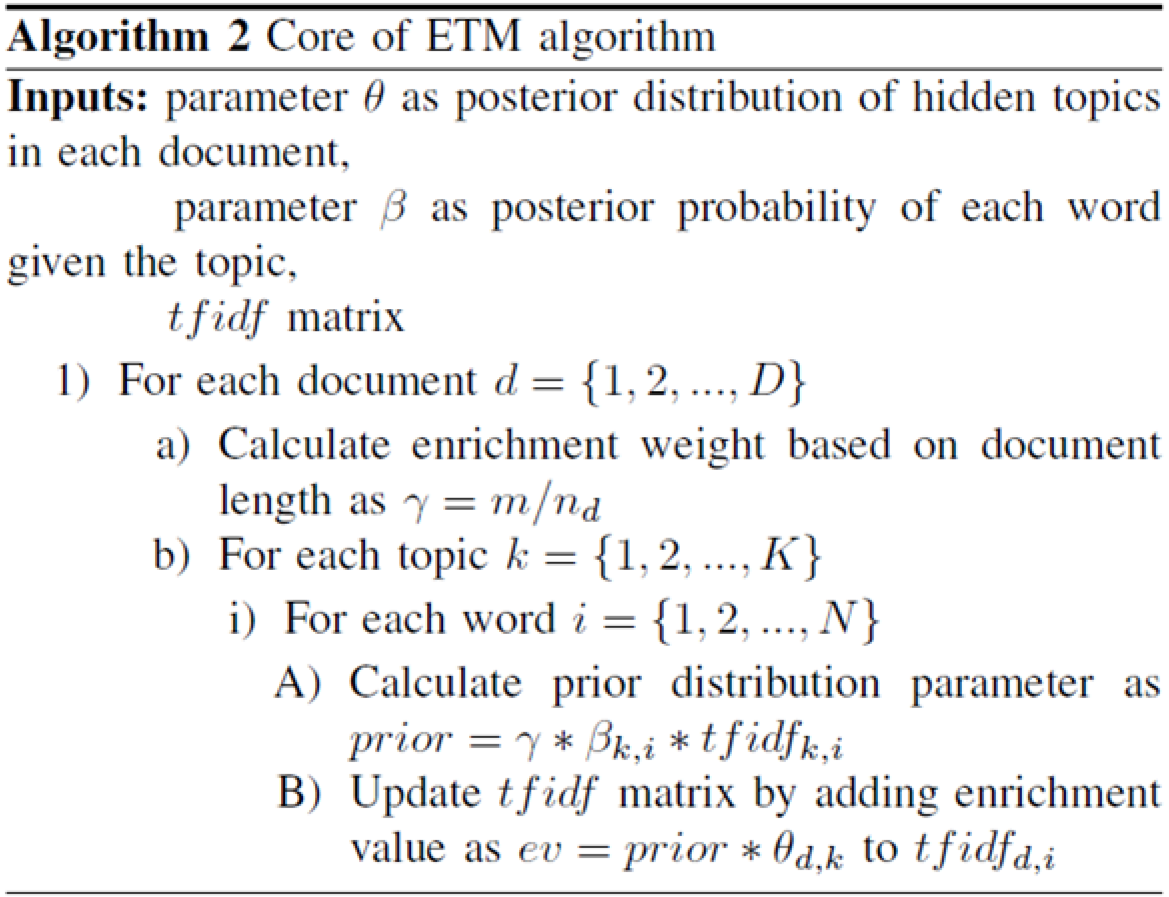




Algorithm 1 LDA generative process used in ETM

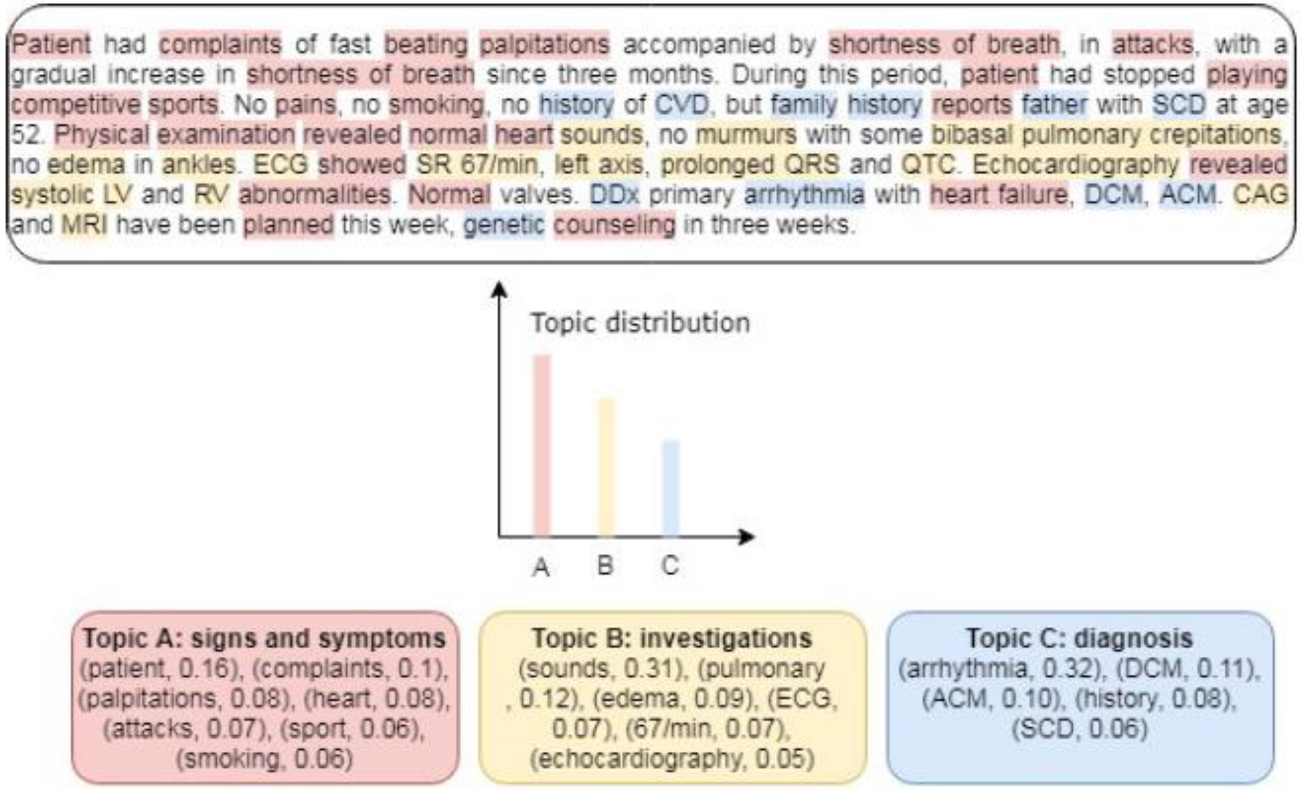
```
1) For each topic $k = \{1, 2, ..., K\}$
    a) Draw a topic-word distribution over the vocabulary $V$ as $\beta_k \sim Dir(\eta)$
2) For each document $d = \{1, 2, ..., D\}$
    a) Draw a document-topic distribution over topics as $\theta_d \sim Dir(\alpha)$
    b) For each word $w$ in document $d$
        i) Draw a topic assignment as $Z_{d,n} \sim Mult(\theta_d)$, where $Z_{d,n} \in \{1, 2, ..., K\}$
        ii) Draw a word $W_{d,n} \sim Mult(\beta_{Z_{d,n}})$, where $W_{d,n} \in \{1, 2, ..., V\}$
```

ETM algorithm



```
Algorithm 2 Core of ETM algorithm
Inputs: parameter $\theta$ as posterior distribution of hidden topics in each document,
        parameter $\beta$ as posterior probability of each word given the topic,
        $tfidf$ matrix
  1) For each document $d = \{1, 2, ..., D\}$
     a) Calculate enrichment weight based on document length as $\gamma = m/n_d$
     b) For each topic $k = \{1, 2, ..., K\}$
        i) For each word $i = \{1, 2, ..., N\}$
           A) Calculate prior distribution parameter as $prior = \gamma * \beta_{k,i} * tfidf_{k,i}$
           B) Update $tfidf$ matrix by adding enrichment value as $ev = prior * \theta_{d,k}$ to $tfidf_{d,i}$
```

ETM example



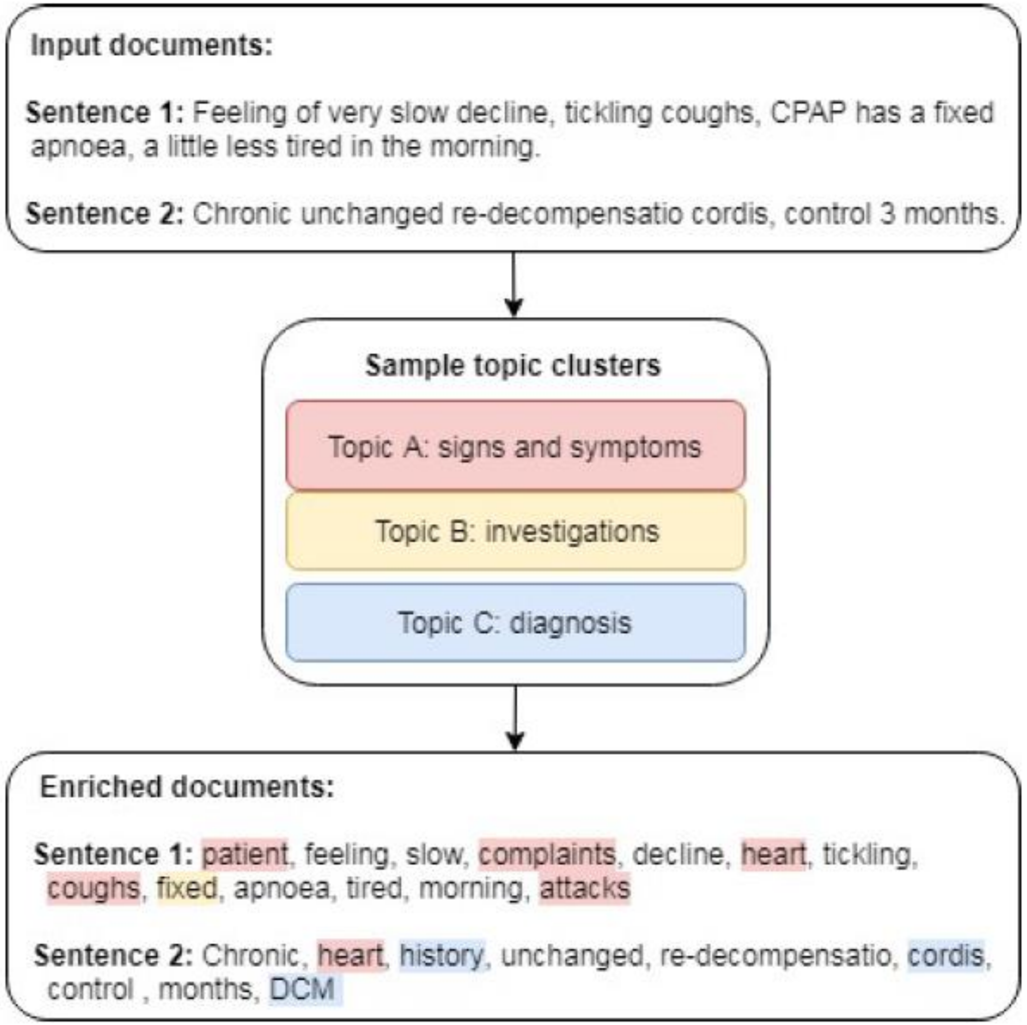
Patient had complaints of fast beating palpitations accompanied by shortness of breath, in attacks, with a gradual increase in shortness of breath since three months. During this period, patient had stopped playing competitive sports. No pains, no smoking, no history of CVD, but family history reports father with SCD at age 52. Physical examination revealed normal heart sounds, no murmurs with some bibasal pulmonary crepitations, no edema in ankles. ECG showed SR 67/min, left axis, prolonged QRS and QTC. Echocardiography revealed systolic LV and RV abnormalities. Normal valves. DDx primary arrhythmia with heart failure, DCM, ACM. CAG and MRI have been planned this week, genetic counseling in three weeks.

Topic A: signs and symptoms
(patient, 0.16), (complaints, 0.1), (palpitations, 0.08), (heart, 0.08), (attacks, 0.07), (sport, 0.06), (smoking, 0.06)

Topic B: investigations
(sounds, 0.31), (pulmonary , 0.12), (edema, 0.09), (ECG, 0.07), (67/min, 0.07), (echocardiography, 0.05)

Topic C: diagnosis
(arrhythmia, 0.32), (DCM, 0.11), (ACM, 0.10), (history, 0.08), (SCD, 0.06)

ETM example Cont.



Input documents:

Sentence 1: Feeling of very slow decline, tickling coughs, CPAP has a fixed apnoea, a little less tired in the morning.

Sentence 2: Chronic unchanged re-decompensatio cordis, control 3 months.

Sample topic clusters

Topic A: signs and symptoms

Topic B: investigations

Topic C: diagnosis

Enriched documents:

Sentence 1: patient, feeling, slow, complaints, decline, heart, tickling, coughs, fixed, apnoea, tired, morning, attacks

Sentence 2: Chronic, heart, history, unchanged, re-decompensatio, cordis, control , months, DCM

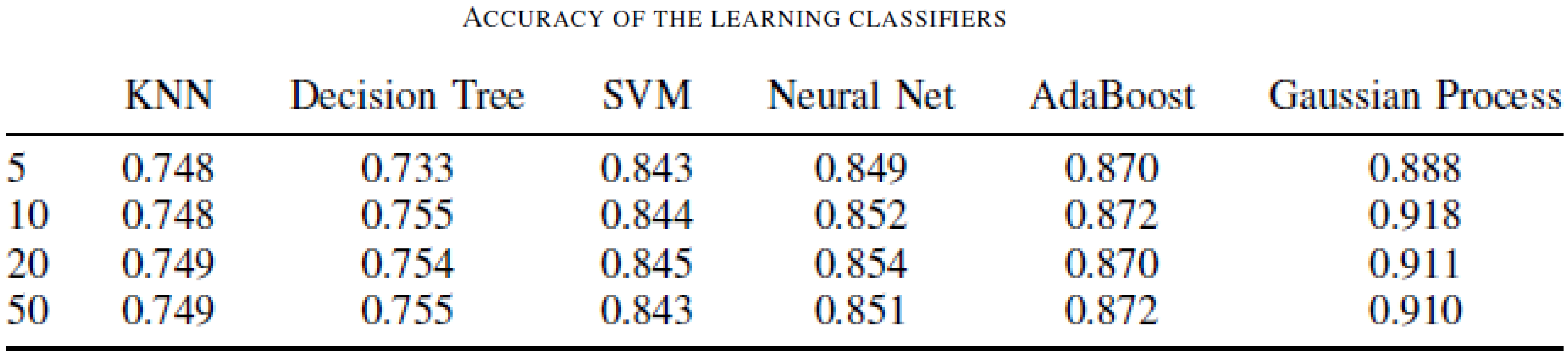
Experiments

Data

- UMCU EHR – Unravel dataset
- Clinical notes from medical doctors or physician assistants
  - Between 2014 and 2018.
- 1002 Dutch clinical notes, manually annotated for medical history
- In total 11,053 sentences:
  - Where 3,560 of them are related to the medical history.
- 20,200 unlabeled clinical cardiovascular sentences have been used for unsupervised topic modeling.

Number of topic clusters

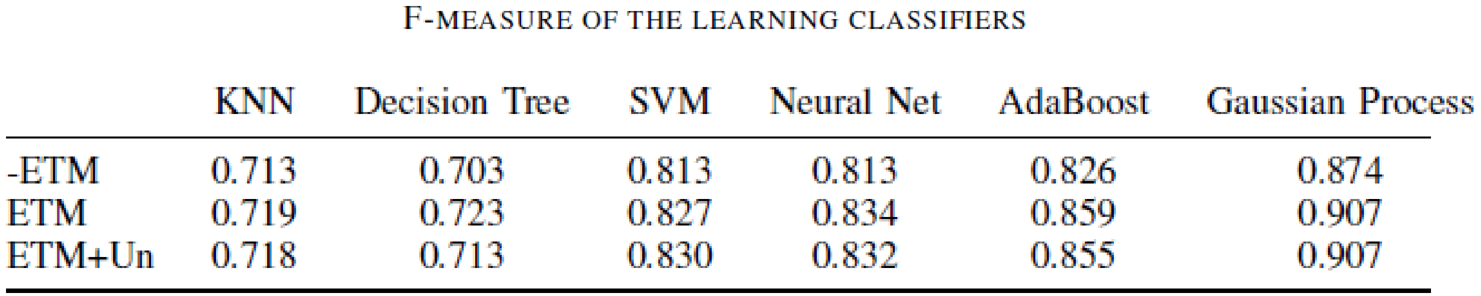
- Best performance is gained by using 10 different clusters



ACCURACY OF THE LEARNING CLASSIFIERS

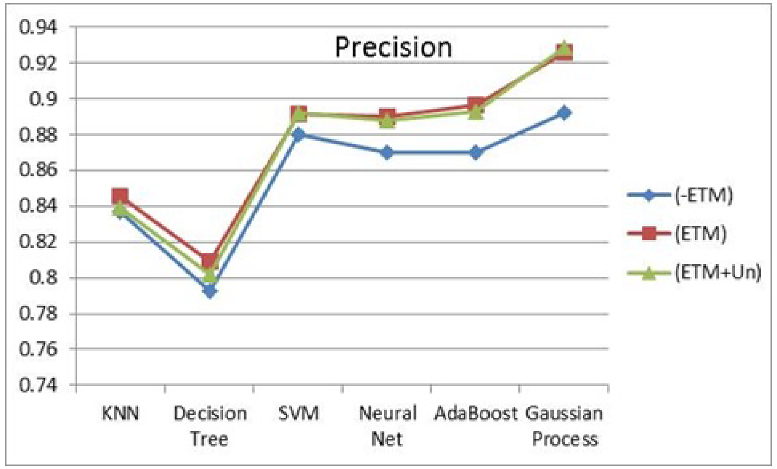
	KNN	Decision Tree	SVM	Neural Net	AdaBoost	Gaussian Process
5	0.748	0.733	0.843	0.849	0.870	0.888
10	0.748	0.755	0.844	0.852	0.872	0.918
20	0.749	0.754	0.845	0.854	0.870	0.911
50	0.749	0.755	0.843	0.851	0.872	0.910

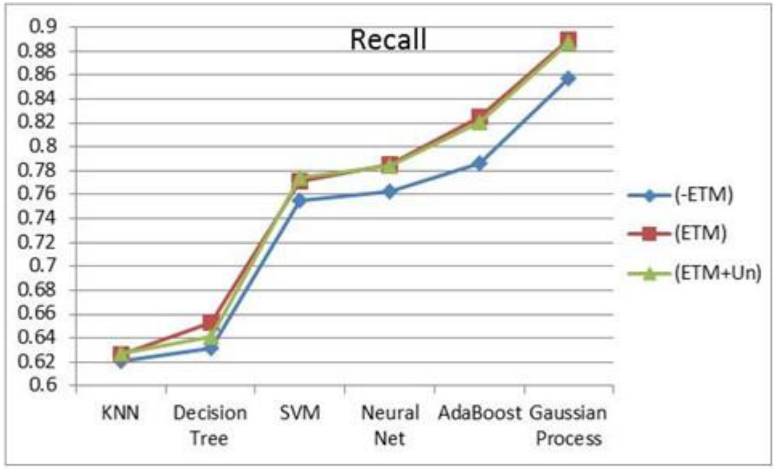
Evaluation of text enrichment



F-MEASURE OF THE LEARNING CLASSIFIERS

	KNN	Decision Tree	SVM	Neural Net	AdaBoost	Gaussian Process
-ETM	0.713	0.703	0.813	0.813	0.826	0.874
ETM	0.719	0.723	0.827	0.834	0.859	0.907
ETM+Un	0.718	0.713	0.830	0.832	0.855	0.907





Summary

- Goal of NLP and text mining:

  “Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.”

- Text mining is needed in medical domain!
- Unsupervised learning can help!

Thank you