Text Analysis of Scientific Publications

TEXT MINING PROJECT 2020

PRESENTED BY:

AMANDA AGYEIWAAH

BHAUMI PANCHAL

UMME RABAB

Index

Motivation

- There has been an exponential growth of information on the internet.
- Most of which is uncategorized and scattered.
- Finding relevant scientific work and publications in a particular field is a cumbersome task.
- Thus our work focuses on categorizing scientific documents depending on their similarities.

Ref: https://dl.acm.org/doi/abs/10.1145/1774088.1774377

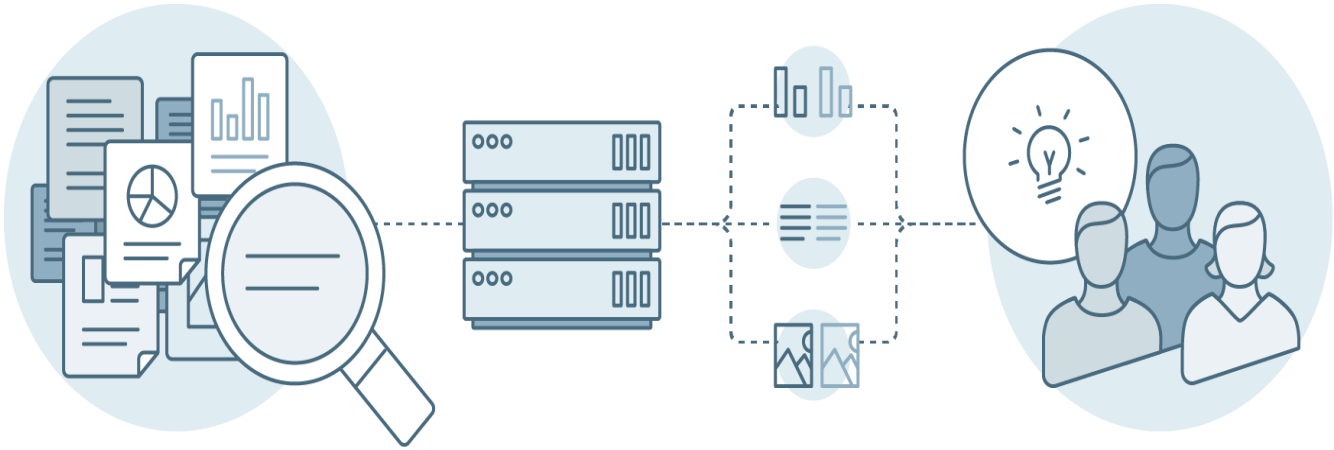
Introduction

- **What is Text Analysis**?

- Search large text quickly.

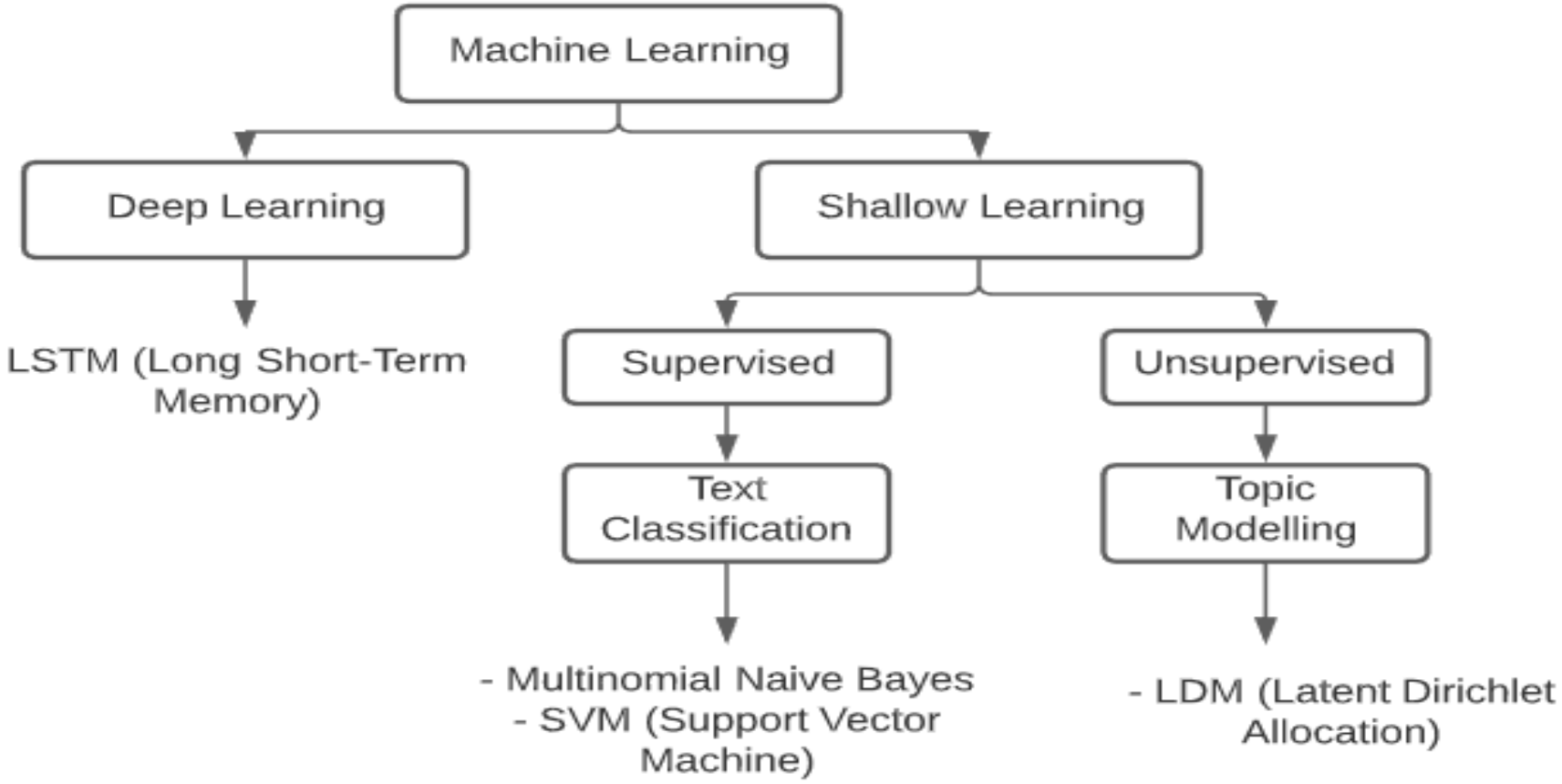
- Conduct complex searches.

- Present the results in a way that suits the study of texts.



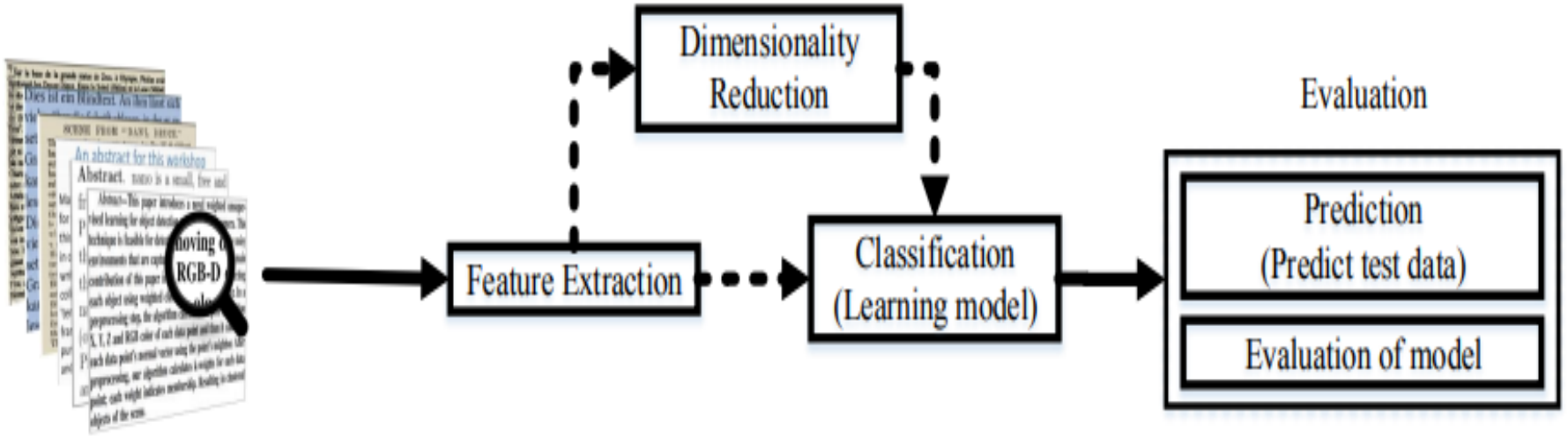
Ref: http://media.springernature.com/full/springer-cms/rest/v1/content/17284494/data/v1

Methodologies



Text Classification

TEXT CLASSIFICATION



Ref: https://www.mdpi.com/2078-2489/10/4/150/pdf

IMPLEMENTATION:TEST CLASSIFICATION

- 443 manually labelled data from ACL Anthropology Corpus was read
- Top words are extracted to form vocabulary.
- First 250 words were considered where there were no abstracts
- Data is converted and loaded into memory ready to feed the model.
- Data Cleaning(removing stop words ,punctuations etc.) was performed
- Data was split in the ratio 70:30 with 310 training & 133 testing
- Data was reshaped to fit the models LSTM, Multinomial Naïve Bayes and SVM.

LABEL DISTRIBUTION

Speech	121
Translation	93
IE	67
Grammar	55
Lexical	35
QA	27
Summarization	23
NE	22

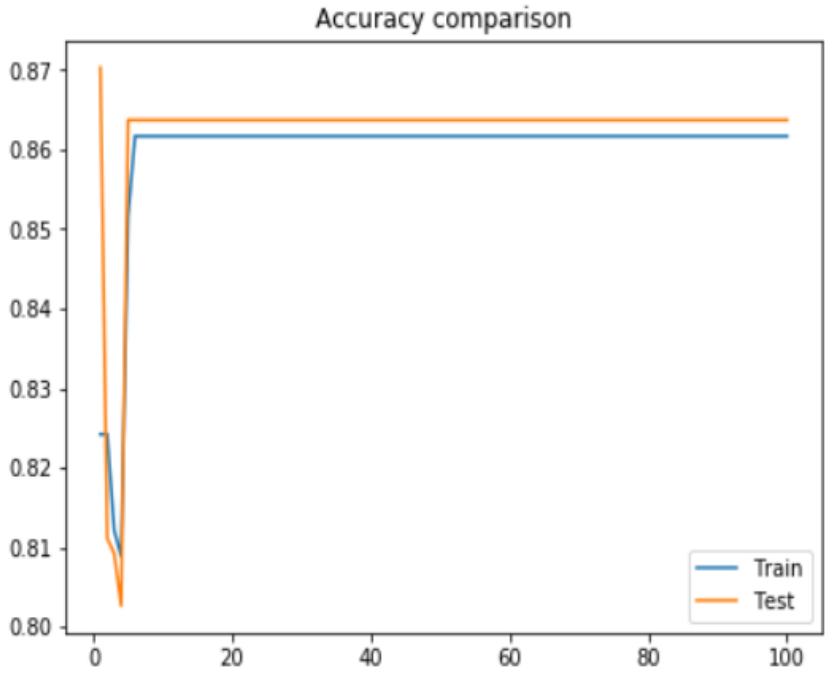
LSTM : Long Short-Term Memory

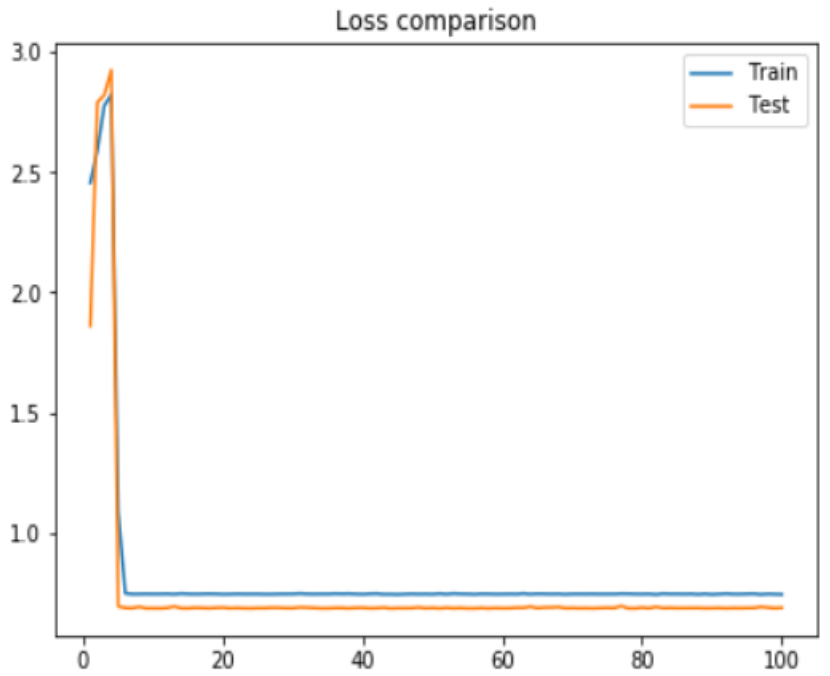
- Deep learning algorithm

•It uses sequential information.

•It keeps in memory previous calculations and uses for prediction of next events.

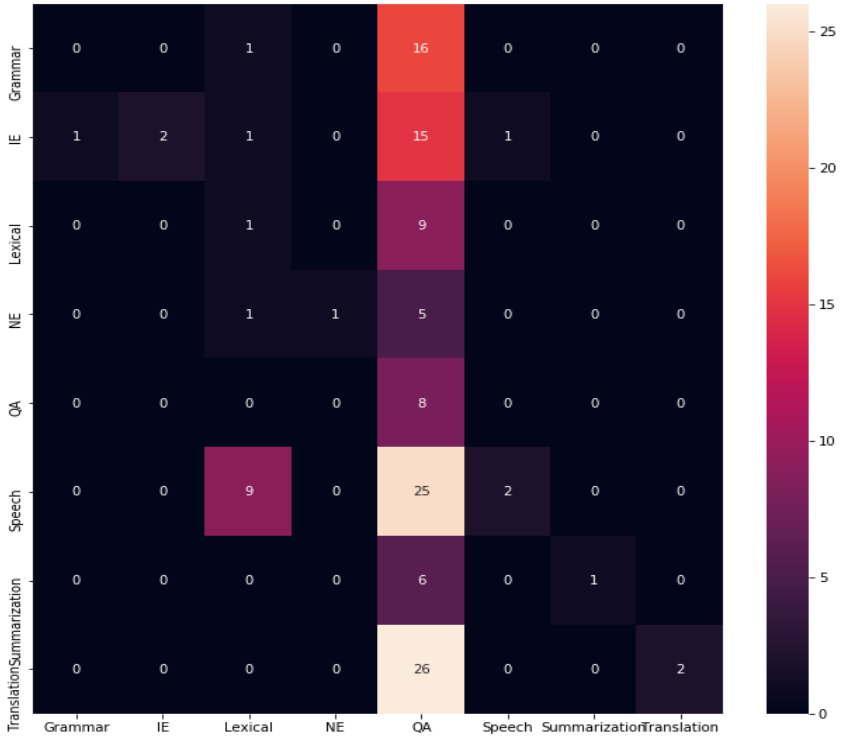
•It is ideal for text analyzing.

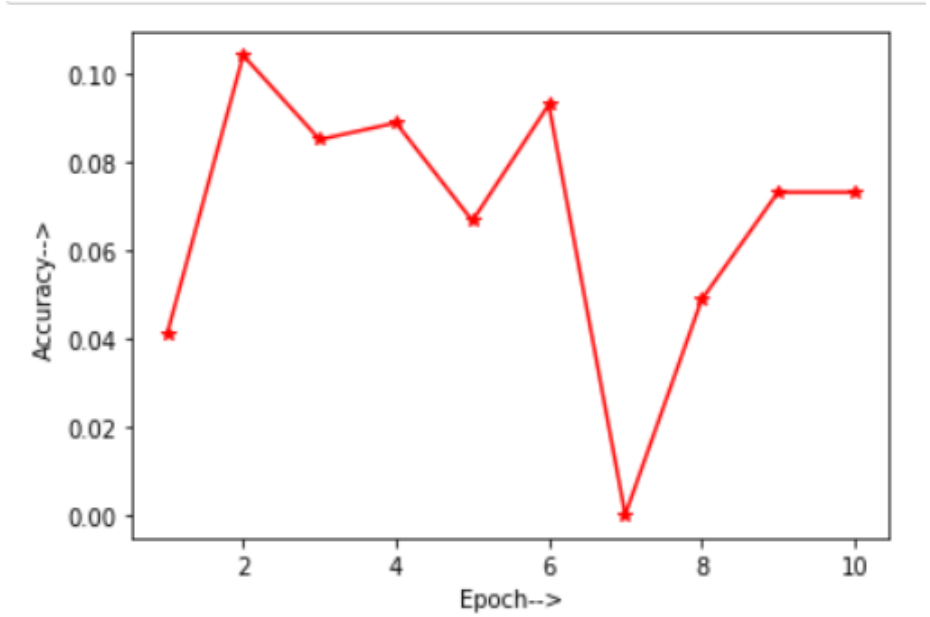




Multinomial Naïve Bayes

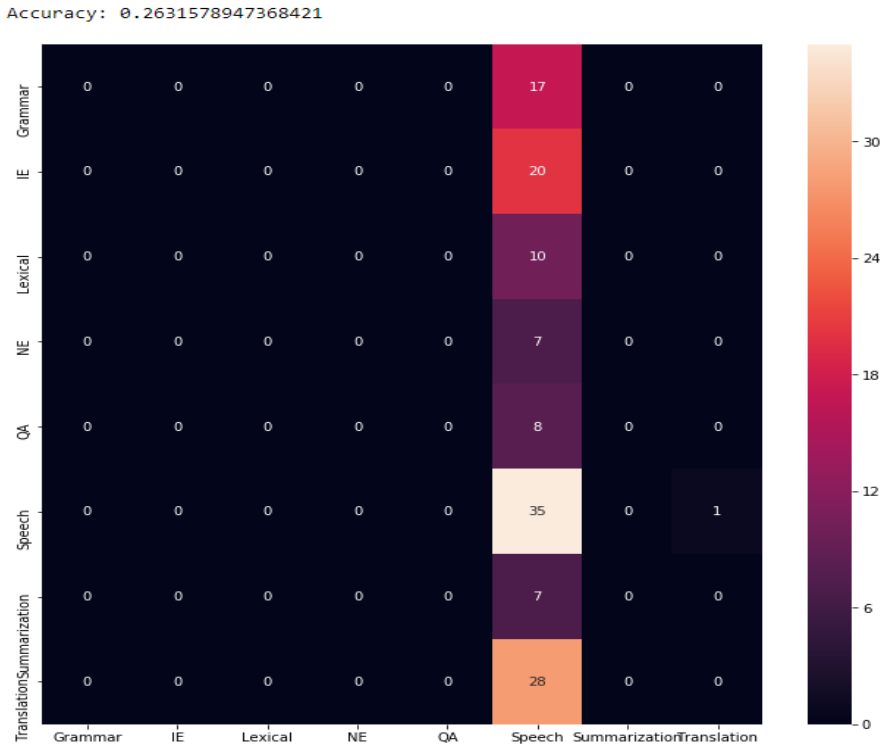
- Useful for probability check and predictions
- Best for small sample data

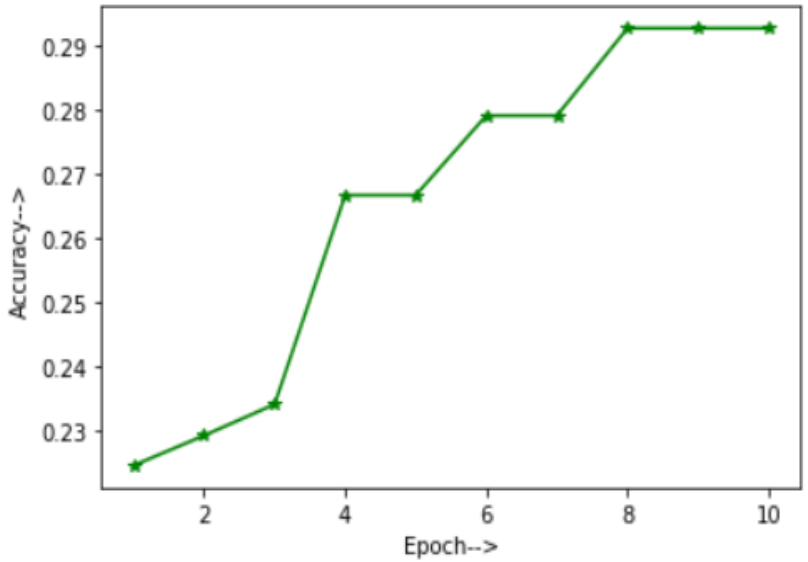




SVM (Support Vector Machine)

- Useful for making best decision boundaries
- Widely used for text classification problems





RESULTS

MODELS	ACCURACY	COMMENTS
LSTM	**0.865**	Performed best
Multinomial Naïve Bayes	**0.15**	Worst Performer
Support Vector Machine	**0.33**	More was expected

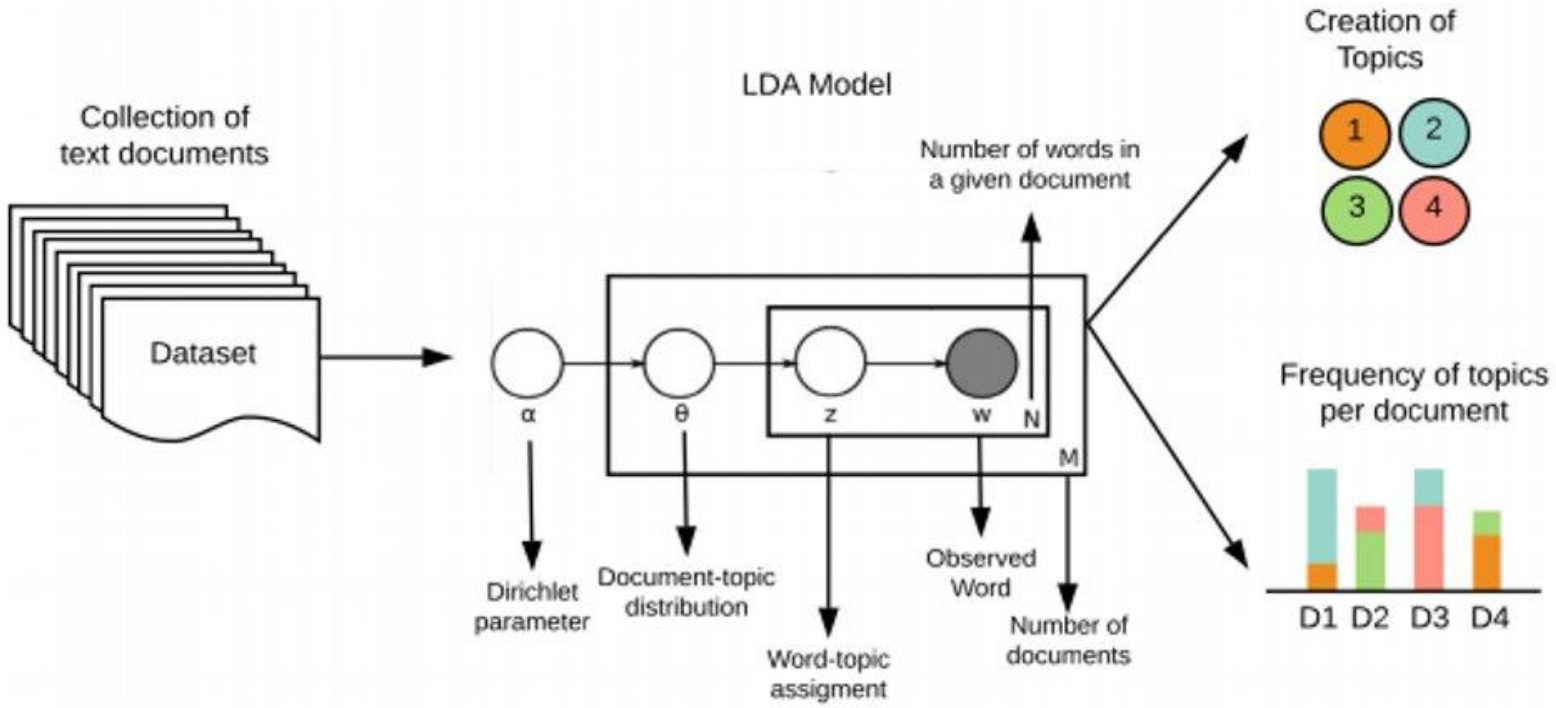
Topic Modelling

Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) - Generative probabilistic model

- Typically used to detect underlying topics in texts.
- It assumes that :
  - Documents are probability distributions over topics.
  - Topics are probability distributions over words.
- Parameters :
  - α - Document-topic density
  - β - Topic-word density
  - Nr. of topics - Guessing amount of topics present in a corpus
  - Nr. of passes - Iterations

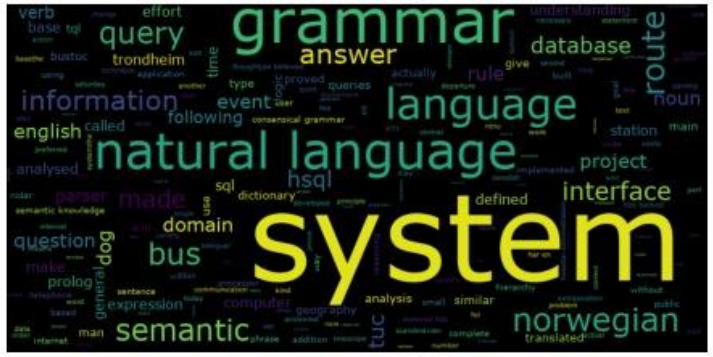
Latent Dirichlet Allocation (LDA)



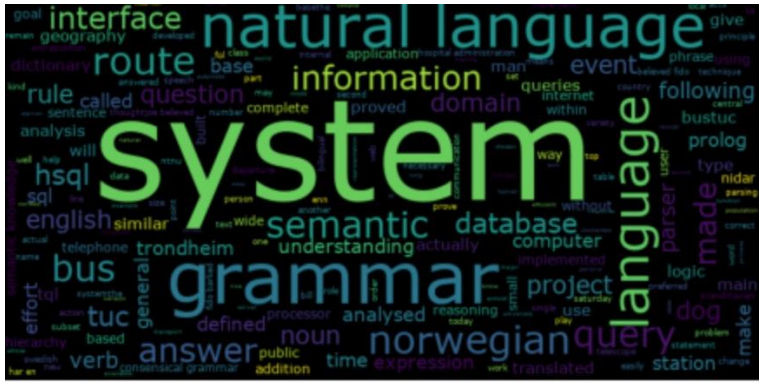
https://arxiv.org/abs/1607.00653v1

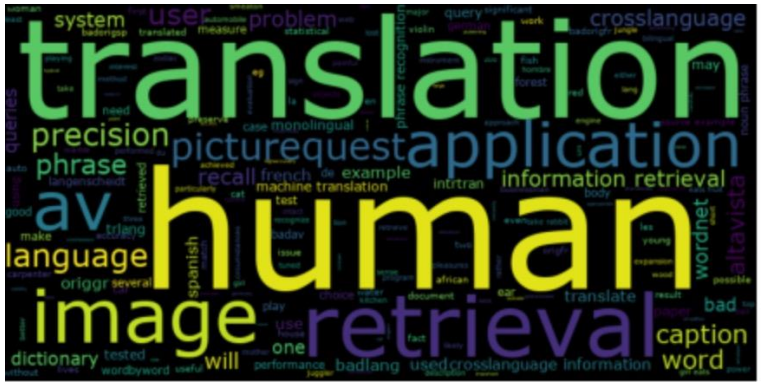
Implementation I

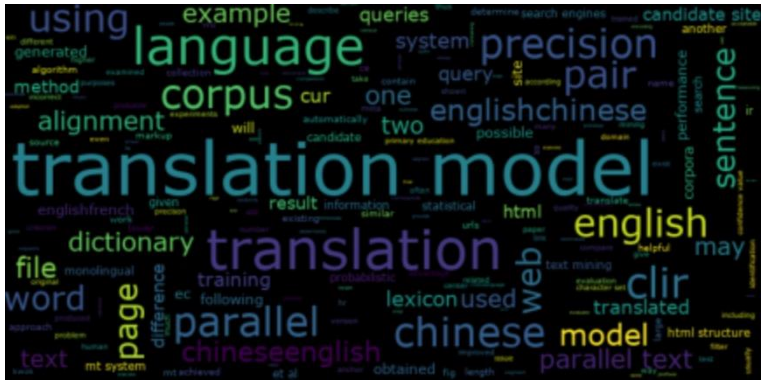
- Data gathering -> Data cleaning -> Exploratory data analysis -> LDA
- Data gathering – First 100 documents picked from the ACL Anthology Reference Corpus.
- Data cleaning - Using libraries like re and string, removed punctuations from the text, numbers and square brackets from the text.
- Exploratory data analysis – Generated word clouds of each document as well as combined :



Implementation II







https://arxiv.org/abs/1607.00653v1

Implementation III - LDA

```
lda = models.LdaModel(corpus = corpus, id2word = id2word, num_topics = 5,
passes = 5)
lda.print_topics()
```

LDA - Results

```
[(0,
  '0.007*"representation" + 0.005*"structures" + 0.005*"module" + 0.005*"dialogue" + 0.005*"structure" + 0.004*"use
r" + 0.004*"generation" + 0.004*"rules" + 0.004*"framework" + 0.004*"used"'),
 (1,
  '0.008*"dialogue" + 0.006*"information" + 0.006*"user" + 0.005*"event" + 0.005*"systems" + 0.005*"dialogues" + 0.
004*"users" + 0.004*"human" + 0.004*"types" + 0.004*"events"'),
 (2,
  '0.013*"dialogue" + 0.008*"translation" + 0.007*"information" + 0.007*"text" + 0.007*"sentence" + 0.006*"target"
+ 0.006*"word" + 0.006*"source" + 0.005*"transfer" + 0.005*"example"'),
 (3,
  '0.013*"translation" + 0.010*"language" + 0.005*"grammar" + 0.004*"javox" + 0.004*"application" + 0.004*"human" +
0.004*"example" + 0.004*"dictionary" + 0.004*"information" + 0.004*"model"'),
 (4,
  '0.032*"coreference" + 0.012*"romanian" + 0.010*"english" + 0.009*"heuristics" + 0.007*"resolution" + 0.007*"core
f" + 0.006*"heuristic" + 0.006*"swizzle" + 0.006*"data" + 0.006*"cocktail"')]
```

https://arxiv.org/abs/1607.00653v1

Conclusion

- LSTM model performed better than SVM and Multinomial Naïve Bayes.
- More was expected from SVM but it could not give any decision boundaries.
- There is more scope for data cleaning to achieve better results.
- More labelled data should be considered in training for better results.
- Regarding LDA, increasing the number of topics and passes stabilize the topic distribution.
- Distribution should be visualized to determine which document belongs to which topic

Thank you

Questions ?