A Distributed Arabic Text Classification Approach Using Latent Semantic Analysis for Big Data

Hadeel Alazzam
Department of Computer Science
University of Jordan

Abdulsalam Alsmady
Department of Computer Engineering
Jordan University of Science and Technology

Outline

- Abstract
- Introduction
- Text Classification
- Latent Semantic Analysis
  - ✓ Sigular Value Decomposition
- Distributed Arabic Text Classification Approach
  - ✓ Training (Learning) Phase
  - ✓ Testing Phase
- Results and Discussion
- References

Abstract

- Big data have special concerns from researchers
- Latent Semantic Analysis (LSA) deals with the semantics of the words.
- LSA has an effective performance in classification
- A distributed text classification approach based on LSA and Cosine Similarity proposed in this paper
- The proposed approach can be applied on big data

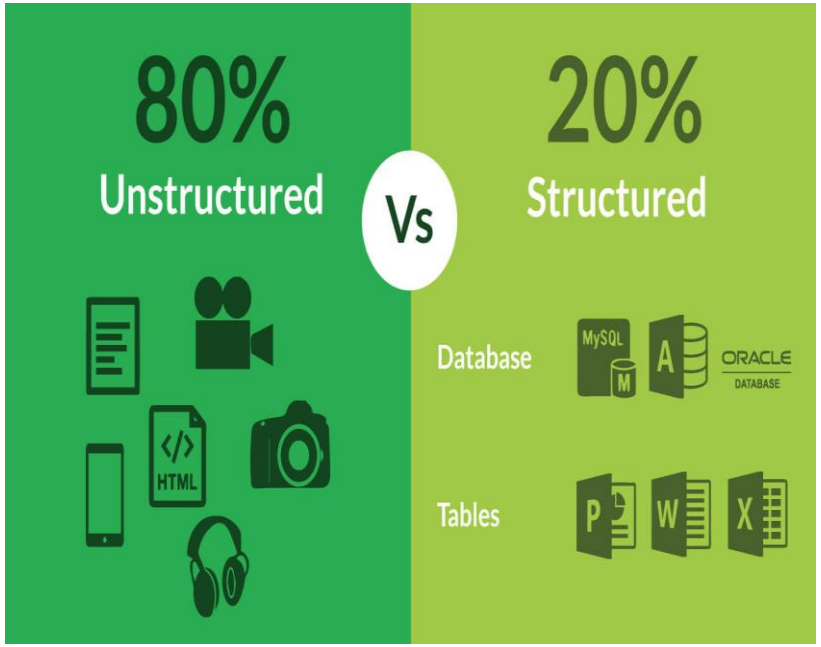
Introduction

- The term big data refer to the large amount of data that can be characterized by the 3V; Volume, Velocity, and Variety
- Big data contains valuable information that can be used in decision making, finance, marketing, etc.
- Many techniques have been proposed to extract valuable information from big data
- Text classification is one of the problems that has been widely studied in information retrieval, data mining, database, and machine learning

Text Classification

- The aim of the text classification is assigning an unlabeled set of documents into predefined categories in an automatic manner
- the process started with a set of documents (training set) that are labeled with class's l∈L manually
- Text classification has been solved using several techniques; vector space model, Naïve Bayes, support vector machine, Neural Network, etc.
- LSA used with a cosine similarity measure for text classification, gives better results than other techniques, but:
  - LSA requires intensive time complexity and computational memory

Big Data



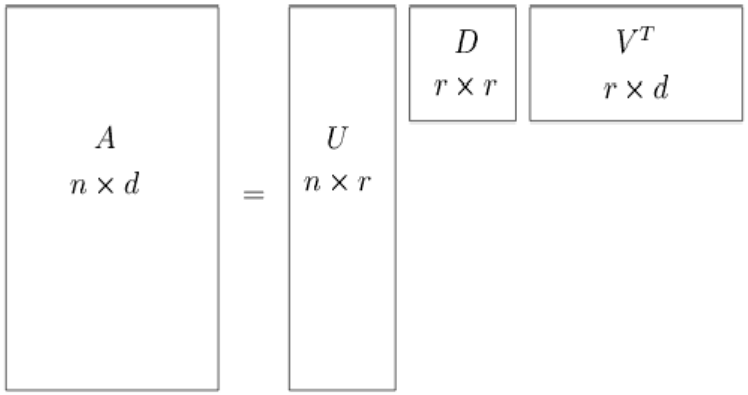
If the value gained from the structured data is useful, then there is a massive potential value gained from unstructured data

LSA

- A technique used for representing documents as a vector
  - ✓ Helps in finding similarities between documents by calculating the distance between vectors
- LSA represent a text as a matrix
  - ✓ The rows represent the terms
  - ✓ The columns represent the documents
  - ✓ The intersection between term and document have a value calculated using Term Frequency-Inverse Document Frequency (TF-IDF)
- TF-IDF reflect how important a word regarding a document

Singular Value Decomposition (SVD)

- LSA uses SVD to refine the Term Document Matrix (TDM)
- The SVD decomposes the TDM into three matrices in order to emphasize the strong relations between words and documents, and remove the noise
- The decomposition expose all the important properties and features of the matrices



Distributed Arabic Text Classification Approach

- The proposed distributed classification approach consist of two phases; learning phase, and testing phase
- **Learning phase:**
  - ✓ A sample of documents is selected randomly from a set of documents, and classified manually
  - ✓ The manually classified sample are distributed among machines
  - ✓ Every machine takes only a single class in order to speed up the process

Learning Phase

- Each machine is supposed to do the following:
  - ✓ Preprocessing the data by removing stop words, normalization, and stemming.
  - ✓ Next, the documents are converted into vectors.
  - ✓ Apply SVD on the TDM, the result of this step will be terms matrix.
  - ✓ Each machine sends the corresponding terms matrix, and documents matrix to all other machines.
- Now, all the machines have the results of the SVD of each class

Testing phase

- Select a documents from the set of the manually classified documents
- compare the results returned from the proposed system with manually results
- Every document is converted to a vector by the following equation

$$\vec{d} = \frac{\vec{t_1} + \vec{t_2} + \cdots + \vec{t_n}}{n}$$

Where is the document vector, and is the term vector from the TDM.

Testing phase ...

- Cosine similarity is used to find relativity between the document and each class
  - ✓ The document will be classified to the class that has the highest value of cosine similarity
  - ✓ If the results are no satisfying, then the number of documents used in training phase must be increased

$$Cosine\ Similarity = \frac{A.B}{||A||||B||}$$

Results and Discussion

- The dataset contains 4000 documents for 10 different categories
- The training set contains 400 documents, 40 document for each category
- The proposed distributed method compared with the work done by [11] and evaluated in terms of precision, recall, and time complexity

Precision Results

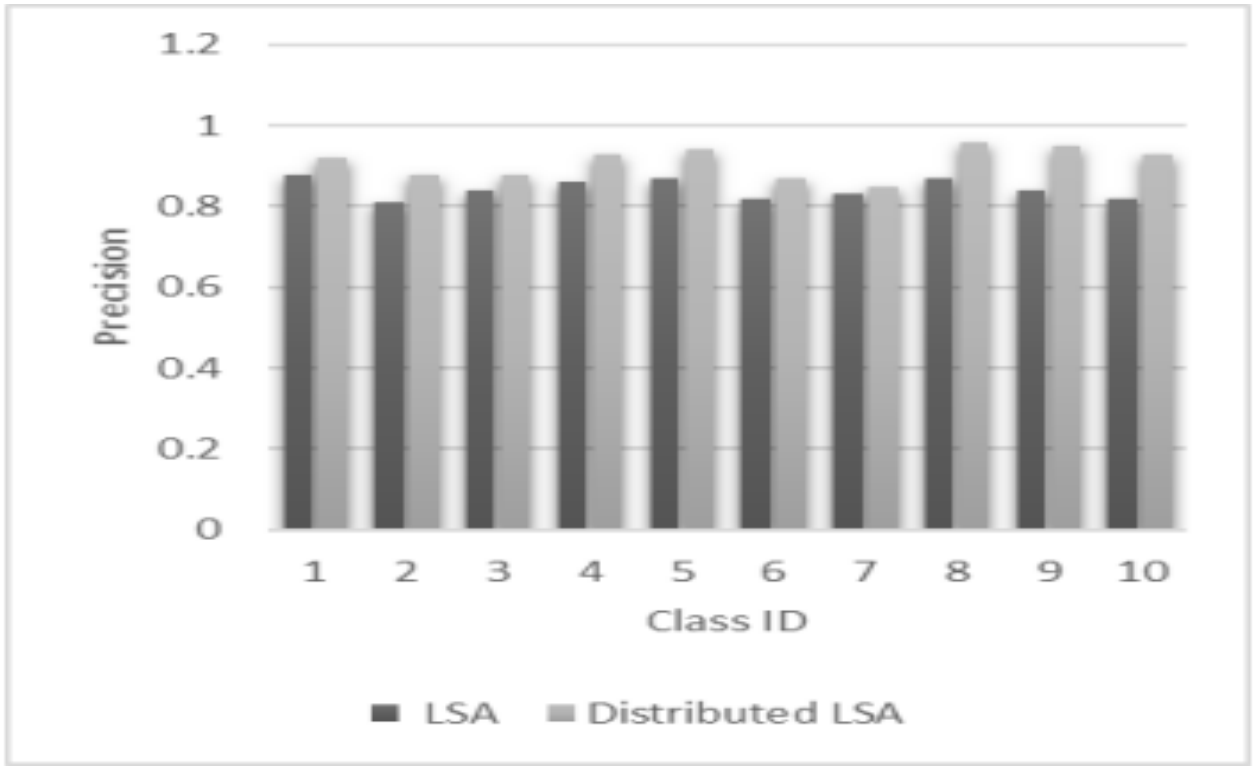


Fig 1: Precision Results for each class

Recall Results

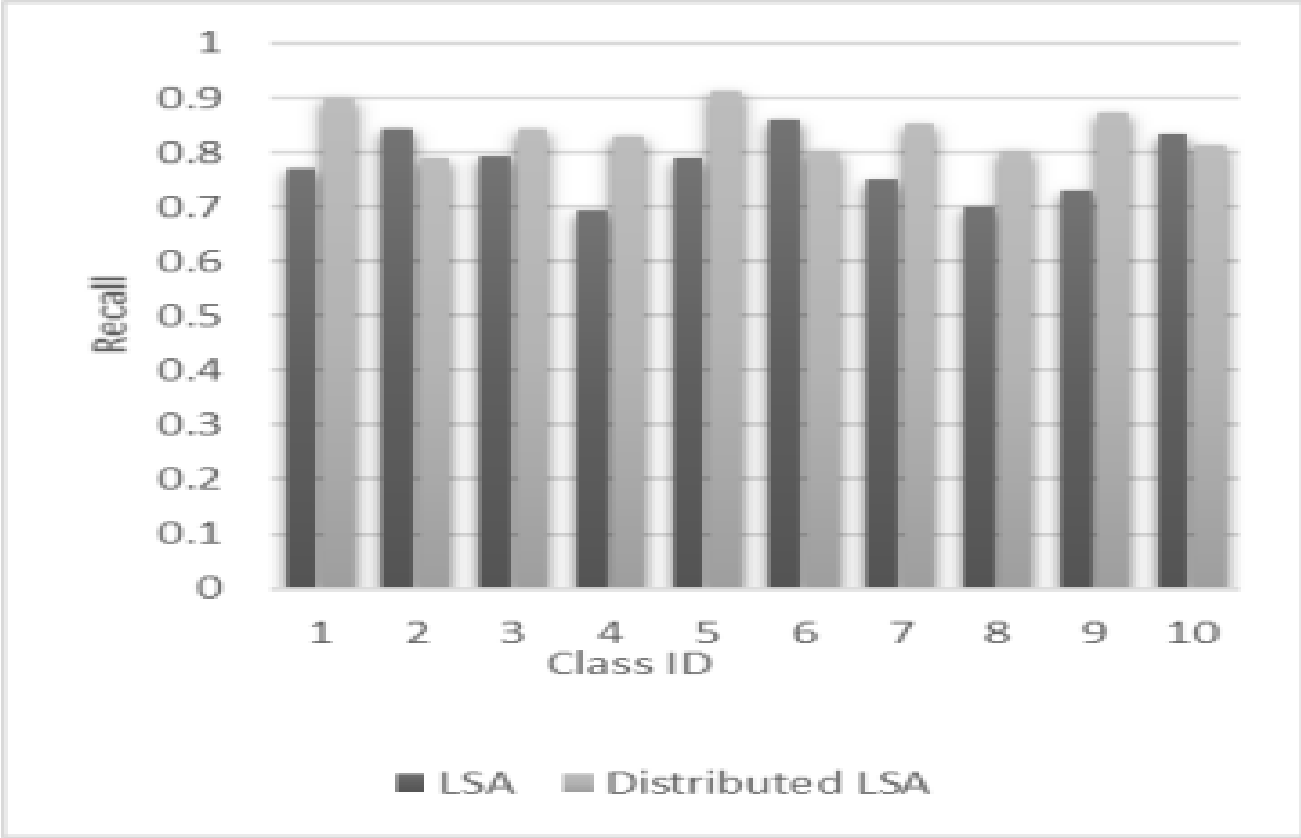


Fig 2: Recall Results for each class

Time Complexity Results

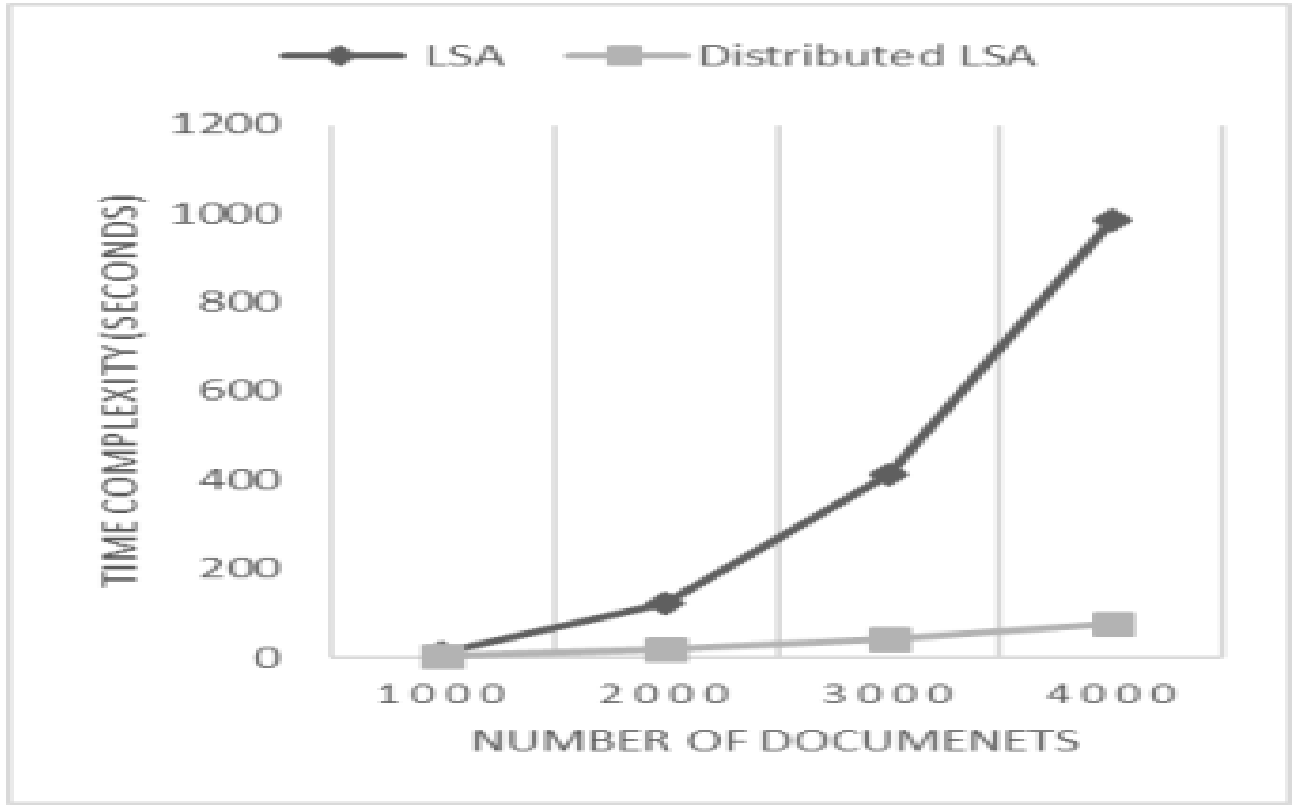


Fig 3: Time Complexity for different datasets

Conclusion

- LSA is a powerful technique for text classification
- A distributed classification method using LSA and cosine similarity proposed based on the method proposed in related works
- The proposed approach aims to reduce the time complexity of the LSA classifier in order to deal with big data
- The proposed approach reduced the time required for classification, and outperform the related works in terms of precision, and recall

Question?