Evaluating LSA Sensibility to Disclosure in Learners’ Interactions

Thursday, March-17-16
Outlines

• Introduction
• Related Work
• LSA-based method for privacy preserving interactions
• Testing and Validation
• Conclusions and Future Work
• Introduction
• Related Work
• LSA-based method for privacy preserving interactions
• Testing and Validation
• Conclusions and Future Work
INTRODUCTION

- Today’s distant learning environment
  - problem: lack of face-to-face interactions
  - solution: collaborative strategies and social interactions tools (Puustinen et al., 2015)

- Example of collaborative strategies: Peer feedback
  - soliciting co-learners to provide feedback in response to learner request (Pridmore and Overocker, 2014)
  - decrease learners’ social isolation
  - encourage motivation and engagement in learning activities
INTRODUCTION

- Peers feedback in educational context (Zhao et al., 2012)
  - in form of corrections, opinions, suggestions, ideas, etc.
  - different types, mainly two:
    - cognitive
      - context independent
      - targeting work content
      - example: peer assessment in writing
    - affective
      - context dependent
      - using affective language (praising, criticism)
      - targeting individual performance
Example of peer affective feedback posted on English forum discussion:
Example of peer affective feedback posted on English forum discussion:

I AM ORIGINALLY FROM CHINA.
I FEEL VERY FRUSTRATED WHENEVER
I HAVE TO TALK IN ENGLISH... I START STUMMERING.
Example of peer affective feedback posted on English forum discussion:
Example of peer affective feedback posted on English forum discussion:
INTRODUCTION

Learners involved in interaction process (Zhao et al., 2012)

- express themselves freely
- share personal experiences
- disclose information about themselves to others (sometimes unwillingly)

Example of a learner request posted on an English forum discussion:

“I am 22 years old, engineering student from India and my family cannot speak English and I am feeling bad to speak with my American girlfriend because I could be wrong...”
**Introduction**

- Personal information disclosed
  - age: 22 years
  - origin: Indian
  - education qualification: student engineer
  - relationship status: American girlfriend

- Privacy threats, if personal data is exposed, or misused by abusive users (Lee et al., 2013)
  - *psychological damage* (cyber-bullying: origin, race, religion, etc.)
  - *social and financial damage* (identity theft, or impersonation, etc.)
Introduction

- Consequences of personal information disclosure in educational context (Puustinen et al., 2015)
  - unsafe e-learning environment
  - abandon of learning

- What is needed in this context?
  - scrutinize learners’ interactions
  - detect and minimize disclosure of personal data
  - create favorable learning environment
  - protect users from privacy risks
**INTRODUCTION**

- **Existing work**
  - lack of research on affective feedback in educational contexts
  - no solution for self-disclosure risks in educational contexts
  - no solution for disclosure in natural language interactions

- **Natural language tasks in educational context:**
  - Southavilay et al., 2013: analysis of collaborative writing processes evolution
  - Nye et al., 2014: evaluation of students’ answers in Intelligent Tutoring System
  - Selmi et al., 2014: semantic analysis for privacy preserving peer feedback
• Introduction
• **Related Work**
  - LSA-based method for privacy preserving interactions
• Testing and Validation
• Conclusions and Future Work
A. Semantic Analysis Tasks and Techniques

• support collaborative activities
  - examine discussed topics between students in discussion forum (Despotakis et al., 2013)
  - visualize and analyse collaborative writing process by extracting semantic topics associated to its evolution (Southavilay et al., 2013)

• automatic evaluation of students’ responses
  - compare responses to predefined model by examining the differences between semantic vectors of responses and model (Nye et al., 2014)

• protect students’ privacy
  - explore semantics of students’ interactions
  - discard negative feedback using Latent Semantic Analysis (Selmi et al., 2014)
B. LSA: SOLUTION FOR SELF-DISCLOSURE

- recognizing personal self disclosed data
- hiding or modifying disclosed data
- stripping sentences revealing personal information

- Latent Semantic Analysis (LSA)
  - technique of vectorial semantics
  - patented in 1988 by Scott Deerwester, Susan Dumais, George Furnas, Richard Harshman, Thomas Landauer, Karen Lochbaum and Lynn Streeter (Deerwester et al., 1990)
  - called Latent Semantic Indexing (LSI) in information retrieval context
**Latent Semantic Analysis**

- Technique for extracting and representing the contextual-usage meaning of words based on statistical computations applied to a large corpus of text

- LSA is used in:
  - data clustering and text classification
  - cross language retrieval
  - text summary
  - questions answering systems

- LSA is based on 3 steps:
  - *occurrence matrix construction*
  - *matrix decomposition*
  - *dimensionality reduction* (low-rank approximation)
First step in LSA: occurrence matrix construction

- Input matrix $A$ representing text of peers’ provided feedback
  - columns: sentences of feedback
  - rows: terms appearing in feedback

$$A = d_j \downarrow \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{bmatrix}$$

$$t_i^T \rightarrow$$
Example:

Request: « Hello. I am from Georgia and I want speak English. Who can help me to practice my English?? Can you for me some advice? »

- Feedback 1: « I would like to practice English with you. Please add me on skype. My skype id is **** »
- Feedback 2: « No pain... no gain »
- Feedback 3: « I will study English with u every day »
**Example:** Occurrence matrix construction

<table>
<thead>
<tr>
<th>Request</th>
<th>Feedback1</th>
<th>Feedback 2</th>
<th>Feedback 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Georgia</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>want</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>speak</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>English</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>help</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>practice</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>like</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>add</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Skype</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>pain</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gain</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>study</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Latent Semantic Analysis

- Second step in LSA: matrix decomposition
  - applying a factorization method called Singular Value Decomposition (SVD) to derive latent semantic structure (Deerwester et al., 1990)
  - decomposing matrix A into 3 matrices

\[ A = U \Sigma V^T \]

- A: input matrix with dimensions \( t \times d \)
- U: \( t \times m \) matrix of extracted topics or concepts (columns)
- \( \Sigma \): \( m \times m \) diagonal matrix containing scaling values sorted in descending order
- V: \( m \times d \) matrix of extracted concepts from the provided feedback (rows)
Latent Semantic Analysis

Evaluating LSA Sensibility to Disclosure in Learners' Interactions
**Example:** Occurrence matrix A

<table>
<thead>
<tr>
<th>Request</th>
<th>Feedback1</th>
<th>Feedback 2</th>
<th>Feedback 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Georgia</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>want</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>speak</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>English</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>help</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>practice</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>like</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>add</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Skype</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>pain</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gain</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>study</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
**Example:** Decomposition of matrix $A$ into 3 matrices $U$, $S$ and $V$

\[
S = \begin{bmatrix}
1. & 0. & 0. \\
0. & 0.6613761 & 0. \\
0. & 0. & 0.5032796 \\
0. & 0. & 0.4571310 \\
0. & 0. & 0. \\
0. & 0. & 0. \\
0. & 0. & 0. \\
0. & 0. & 0. \\
0. & 0. & 0. \\
0. & 0. & 0. \\
0. & 0. & 0. \\
0. & 0. & 0. \\
0. & 0. & 0. \\
0. & 0. & 0.
\end{bmatrix}
\]
Third step in LSA: dimensionality reduction

- select \( k \) greatest singular values to construct an approximation \( A' \) of matrix \( A \)
- **Example:** dimensionality reduction of A with $k=3$

![Singular values diagram](image)
**LATENT SEMANTIC ANALYSIS**

- **Third step in LSA:** dimensionality reduction and construction of the approximation

\[
\begin{align*}
\text{terms} & \quad \text{documents} \\
0.3504278 & - 0.0094777 & 0. & 0.0426393 \\
0.3504278 & - 0.0094777 & 0. & 0.0426393 \\
0.3504278 & - 0.0094777 & 0. & 0.0426393 \\
0.3835894 & 0.4041210 & 0. & 0.1388072 \\
0.3504278 & - 0.0094777 & 0. & 0.0426393 \\
0.3409500 & 0.3298962 & 0. & 0.1168641 \\
- 0.0094777 & 0.3393739 & 0. & 0.0742248 \\
- 0.0094777 & 0.3393739 & 0. & 0.0742248 \\
- 0.0094777 & 0.3393739 & 0. & 0.0742248 \\
4.492D-19 & 3.612D-17 & 0.3558724 & - 7.045D-17 \\
4.492D-19 & 3.612D-17 & 0.3558724 & - 7.045D-17 \\
0.0426393 & 0.0742248 & 0. & 0.0219431
\end{align*}
\]
OUTLINE

• Introduction
• Related Work
• LSA-based method for privacy preserving interactions
  • Mining Step
  • Composition Step
• Test and Validation
• Conclusions and Future Work
**OUR APPROACH**

- **Goals of this work**
  - *scrutinize* learners’ interactions
  - *detect* and *minimize* disclosure of personal data
  - create *favorable* learning environment
  - *protect* users from privacy risks

- **2-step proposed approach**
  - *mining step*
    - discarding negative feedback messages that negatively affect learning
  - *composition step*
    - eliminating any self-disclosing sentences from mined feedback
    - reconstructing new feedback
ARCHITECTURE

Fig. 1. Architecture of our approach
OUR APPROACH

- Goals of this work
  - scrutinize learners’ interactions
  - detect and minimize disclosure of personal data
  - create favorable learning environment
  - protect users from privacy risks

- 2-step proposed approach
  - mining step
    - discarding negative feedback messages that negatively affect learning
  - composition step
    - eliminating any self-disclosing sentences from mined feedback
    - reconstructing new feedback
Architecture of our approach

Fig. 1. Architecture of our approach
OUR APPROACH: MINING STEP

- **Role:** discarding negative feedback affecting learning process
  - negative feedback: bullying, demeaning, or other negative comments

- **Preprocessing:** recognizing negative feedback
  - short text document
  - vector of text attribute values (frequency computing)
  - natural language processing techniques
    - *bag of words* as linguistic model
    - no stop words and non-content bearing words
    - no grammatical structural characteristics or positional information
Example of feedback message

“Whenever I speak to native (English speaker), I feel very frustrated and I'll start to stammer. The phrasing, sentence structure & grammar of my sentences become all in a mess”

Preprocessing steps

* tokenization
  <Whenever, I, speak, to, native, English, speaker, I, feel, very, frustrated, and, I’ll, start, to, stammer, The, phrasing, sentence, structure, & grammar, of, my, sentences, become, all, in, a, mess>

* stop words removal
  <Speak, native, English, speaker, feel, frustrated, start, stammer, phrasing, sentence, structure, grammar, sentences, become, mess>
Example of feedback message

“Whenever i speak to native (English speaker), I feel very frustrated and i'll start to stammer. The phrasing, sentence structure & grammar of my sentences become all in a mess”

Preprocessing steps

• **stemming**: converting token to its morphological format
  
  \[<\text{speak}, \text{speaker}, \text{speaking}> \rightarrow \text{speak}>\]

• **frequency computing**: result of preprocessing steps
  
  \[<\text{Speak}, \text{native}, \text{English}, \text{speaker}, \text{feel}, \text{frustrated}, \text{start}, \text{stammer}, \text{phrasing}, \text{sentence (2)}, \text{structure}, \text{grammar}, \text{become}, \text{mess}>\]
Our Approach: Mining Step

- **Sentiment analysis**: classifying feedback as positive and negative
  - Naive Bayes classifier
    - probabilistic classifier based on Bayes' theorem
    - independence assumptions on words’ position in text (Pang et al., 2002)
    - for a given set of classes, probability of a class:

\[
P(c|d) = P(c) \prod_{1 \leq k \leq n_d} P(t_k|c) \quad (1)
\]

- $c$: target class
- $d$: current document
- $t$: current term
- $n$: number of terms in current document
Sentiment analysis: classifying feedback as positive and negative

- Naive Bayes classifier
  - result: class with the highest probability given the feedback
  - estimation of log probability, given by:

\[
\arg \max_c \log(\hat{P}(c)) \sum_{1 \leq k \leq n_d} \log(\hat{P}(t_k | c)) \quad (2)
\]

- \(c\): target class
- \(t\): current term
- \(n\): number of terms in current document
Fig. 1. Architecture of our approach
OUR APPROACH: COMPOSITION STEP

- **Role**: removing any self-disclosure of personal data in interactions
  - detect self-disclosing messages
  - preventing message *observer* from gaining knowledge of disclosed personal information

- **LSA for privacy protection from self-disclosure in interactions**: 
  - learner message: set of *sentences* and *concepts*
  - sentence: set of *terms*
  - most representative sentences of current concept: set of *terms* best representing that concept
OUR APPROACH: COMPOSITION STEP

- LSA in composition step: we propose to consider
  - three main steps:
    1. matrix creation
    2. SVD application
    3. sentences selection
  - three parameters:
    - weighting schemes
    - approximation rank or number of dimensions
    - similarity measure
Fig. 1. Architecture of our approach
First step in LSA

- input matrix representing input text message
  - columns: sentences of input text
  - rows: terms appearing in message, extracted using preprocessing tools

- weighting schemes parameter
  - affecting differently LSA performance depending on data size and nature
  - three weighting categories: local (Term Frequency), global (Inverse Document frequency) and hybrid (Log Entropy, TF-IDF)

**Hypothesis 1:**
Local weighting schemes are the most appropriate to discard disclosing terms and sentences
**OUR APPROACH: COMPOSITION STEP**

- **Example:** Occurrence matrix A

<table>
<thead>
<tr>
<th>Request</th>
<th>Feedback1</th>
<th>Feedback 2</th>
<th>Feedback 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Georgia</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>want</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>speak</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>English</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>help</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>practice</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>like</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>add</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Skype</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>pain</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gain</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>study</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Example: TF-IDF weighting computing

t₁ = English, R (Request), F1 (Feedback 1), F2 (Feedback 2) and F3 (Feedback 3)

- Term Frequency (TF)
  - TF (t₁, R) = occurrences of t₁ / terms number in R
  - TF (t₁, R) = 2/24

- Inverse Document Frequency (IDF)
  - IDF (t₁) = log (number of docs in corpus / number of documents containing t₁)
  - IDF (t₁) = log (4/3)
Example: TF-IDF weighting computing

t_1 = English, R (Request), F1 (Feedback 1), F2 (Feedback 2) and F3 (Feedback 3)

TF-IDF computing

- TF-IDF (t_1, R) = \frac{2}{24} \times \log \frac{4}{3} = 0.04
- TF-IDF (t_1, F1) = \frac{1}{18} \times \log \frac{4}{3} = 0.04
- TF-IDF (t_1, F2) = \frac{0}{4} \times \log \frac{4}{3} = 0
- TF-IDF (t_1, F3) = \frac{1}{8} \times \log \frac{4}{3} = 0.06
Fig. 1. Architecture of our approach
OUR APPROACH: COMPOSITION STEP

- Second step in LSA
  - SVD on matrix A to derive latent semantic structure
  - decomposing A into 3 matrices to extract significant terms and sentences

\[ A = U \Sigma V^T \]

- **A**: input matrix with dimensions \( m \times n \)
- **U**: \( m \times n \) matrix of extracted topics or concepts (columns)
- **\( \Sigma \)**: \( n \times n \) diagonal matrix containing scaling values sorted in descending order
- **V**: \( m \times n \) matrix of extracted concepts from the provided feedback (rows)
Second step in LSA

- SVD on matrix $A$ to derive latent semantic structure
- decomposing $A$ into 3 matrices to extract significant terms and sentences

- **dimensionality parameter**
  - reducing dimensions to enhance relationships between two terms or documents
  - different impacts on LSA performance
  - no consensus regarding optimal reduction value (generally used $k=300$)

**Hypothesis 2:**
Dimensionality reduction positively affect LSA ability to discard disclosing terms and sentences
Fig. 1. Architecture of our approach
Our Approach: Composition Step

- Third step in LSA
  - selecting important sentences in learners messages
    - request initiating interactions
    - all learners messages to the same request
  - similarity metrics parameter
    - computing similarity between two semantic vectors
    - different metrics: distance (Euclidean distance and Jaccard) and similarity (cosine)
    - affecting differently LSA outputs depending on data size and nature

Hypothesis 3:
Distance metrics are more appropriate to discard disclosing terms and sentences
**OUR APPROACH: COMPOSITION STEP**

- **Example:**

  **Request:** «Hello. I am from Georgia and I want speak English. Who can help me to practice my English?? Can you for me some advice? »

  ➢ **Feedback 1:** «I would like to practice English with you. Please add me on skype. My skype id is **** »
  ➢ **Feedback 2:** «No pain... no gain »
  ➢ **Feedback 3:** «I will study English with u every day »
Example: similarity computing using Cosine and Jaccard between Request and Feedback 1

- **Cosine**
  \[
  \cos(R, F1) = \frac{R \cdot F1}{||R|| ||F1||} = 0.36
  \]

- **Jaccard**
  \[
  d(R, F1) = \frac{|R \cap F1|}{|R \cup F1|} = 0.25
  \]
EVALUATING LSA SENSIBILITY TO DISCLOSURE IN LEARNERS’ INTERACTIONS

OUTLINE

• Introduction
• Related Work
• LSA-based method for privacy preserving interactions
• Testing and Validation
• Conclusions and Future Work
Corpus

- over 1000 comments threads to build semantic space
- forum discussion for English Second Language (ESL)
- challenge: noisy and informal data

Data Preprocessing

- data preprocessed to reduce dimensionality
  - average of 5 peer feedback received for each request
  - short feedback (with less than 3 words) excluded
- cleansing
  - noise removed: “Hhhhhhhhh gr8” (for great)
  - morphological format to reduce noise and misspelling: “scheduale”, “schedual”, “skedul” → schedul
TESTING AND VALIDATION

- Correlations between human judges’ scores
  - 4 point Likert scale regarding request: 1 = “very bad” ; 4 = “very good”
  - 2 independent human judges: inter-rater agreement using Cohen kappa ($\kappa=0.68$)

- Correlations between human judges and LSA scores
  - N_words: number of words
  - Feed_Req: LSA cosine(feedback, request)
  - Avg_Feed: average LSA cosine(feedback, all peers’ feedback to the request)

Table 1. Correlation between human judges and LSA measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>N_words</td>
<td>0.233</td>
</tr>
<tr>
<td>Feed_Req</td>
<td>0.458</td>
</tr>
<tr>
<td>Avg_Feed</td>
<td>0.464</td>
</tr>
</tbody>
</table>
**Testing and Validation**

- Optimal dimensionality using 2 methods
  - energy: 90% of the information
  - cumulative variance: 60% to 80% of the information
  - tested dimensionality values: 25%, 50%, 70%, 80% and 100%

![Fig. 2. Energy and variance methods](image-url)
Hypothesis 1: Weighting
Local weighting schemes are the most appropriate to discard disclosing terms and sentences

Hypothesis 2: Dimensionality
Dimensionality reduction positively affect LSA ability to discard disclosing terms and sentences

Hypothesis 3: Similarity metrics
Distance metrics are more appropriate to discard disclosing terms and sentences
Hypothesis 1: Weighting
Local weighting schemes are the most appropriate to discard disclosing terms and sentences

Hypothesis 2: Dimensionality
Dimensionality reduction positively affect LSA ability to discard disclosing terms and sentences

Hypothesis 3: Similarity metrics
Distance metrics are more appropriate to discard disclosing terms and sentences
Impact of weighting on dimensionality

- local scheme: Binary Term Frequency
- hybrid schemes: Term Frequency-Inverse Document Frequency (TF-IDF) and Log Entropy
- best performance: no-weight (less variance)

Fig. 3. Impact of weighting on dimensionality
**Testing and Validation**

- Impact of dimensionality and weighting on LSA
  - multiple regression analysis
    - variables: independent (Feed_Req, Avg_Feed), dependent (average human judges score)
    - similarity measure: cosine
    - best performance with Log Entropy and no dimensionality reduction

*Fig. 4.* Correlations with human graders: Interaction between dimensionality and weighting
Hypothesis 1: Weighting
Local weighting schemes are the most appropriate to discard disclosing terms and sentences

Hypothesis 2: Dimensionality
Dimensionality reduction positively affect LSA ability to discard disclosing terms and sentences

Hypothesis 3: Similarity metrics
Distance metrics are more appropriate to discard disclosing terms and sentences
Testing and Validation

- Impact of dimensionality and similarity metrics on LSA
  - correlation with human graders
    - $r = 0.64$, $p < 0.001$
    - similarity metrics: Euclidean distance, Jaccard and Cosine
    - best performance model: no reduction, Euclidean distance and TF-IDF

Fig. 5. Correlations with human graders: Interaction between dimensionality and similarity metrics
TESTING AND VALIDATION

- **Testing on entire corpus**
  - implementation of two models
    - model 1: TF-IDF, no dimensionality reduction, Euclidean distance
    - model 2: TF-IDF, no dimensionality reduction, Cosine similarity
  - testing with binary approach
    - conversion of average human scores and LSA scores
    - score < 3 : self-disclosing message (removed)
    - score > 3 : relevant message
  - percentage of right graded messages
    - model 1: 41% (best performance model)
    - model 2: 39%
• Introduction
• Related Work
• LSA-based method for privacy preserving interactions
• Testing and Validation
• Conclusions and Future Work
CONCLUSIONS

- Interactions in distant learning environment
  - necessary to complete learning
  - privacy threats and challenges associated to self-disclosure (cyber-bullying)

create favorable learning environment and protect learners’ privacy

- Two-step based approach
  - mining step: eliminating negative messages causing psychological harm
  - composition step: detecting and removing self-disclosure of personal data
**CONCLUSIONS**

- **Two-step based approach**
  - composition step: based on Latent Semantic Analysis (LSA)
    - highly parameterized
    - tested parameters: dimensionality reduction, weighting schemes, similarity metrics
  - best performance model
    - human judges correlation $r = 0.64$, $p < 0.001$
    - no reduction, TF-IDF, Euclidean distance

- **Challenges**
  - data preprocessing: peers’ interactions informal, many mistakes, and symbols
  - sensitivity of human raters to self-disclosure
Future Work

Future work

- inclusion of advanced natural language processing techniques
  - enhancing coherence of composed feedback
  - resolving problem of loss of information
- inclusion of regular expressions in our approach
  - specifying terms associated to self-disclosure
  - decrease sensitivity of human raters
- larger set of experiments to demonstrate feasibility of proposed approach in real educational scenarios with large volumes of data
THANK YOU

QUESTIONS?
Evaluating LSA Sensibility to Disclosure in Learners’ Interactions

Authors

- **Mouna Selmi**
  Ph.D Student
  selmimou@iro.umontreal.ca

- **Hicham Hage**
  Assistant Professor
  hhage@ndu.edu.lb

- **Esma Aïmeur**
  Full Professor
  aimeur@iro.umontreal.ca
REFERENCES

PEARSON CORRELATION

- Pearson correlation coefficient
  - measure of the linear correlation between two variables X and Y
  
  \[ r = r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \]
  
- interpretation of Pearson's correlation coefficient

<table>
<thead>
<tr>
<th>Strength of Association</th>
<th>Coefficient, ( r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Small</td>
<td>.1 to .3</td>
</tr>
<tr>
<td>Medium</td>
<td>.3 to .5</td>
</tr>
<tr>
<td>Large</td>
<td>.5 to 1.0</td>
</tr>
</tbody>
</table>
Weighting Schemes (1/2)

- Term frequency (TF)
  - *raw frequency*: number of occurrences of term $t$ in document $d$
  - local approach

- Inverse Document Frequency (IDF)
  - global approach

\[
idf(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}
\]

With
- $N$: total number of documents in the corpus
- $|\{d \in D : t \in d\}|$: number of documents where the term $t$ appears

- Term frequency-Inverse Document frequency (TF-IDF)
  - hybrid approach

\[
tfidf(t, d, D) = tf(t, d) \times idf(t, D)
\]
Weighting Schemes (2/2)

- Term frequency (TF)
  - raw frequency: number of occurrences of term t in document d
  - local approach

- Inverse Document Frequency (IDF)
  - global approach

With
- N: total number of documents in the corpus
- |{d ∈ D : t ∈ d}|: number of documents where the term t appears

- Term frequency-Inverse Document frequency (TF-IDF)
  - hybrid approach
  - assigning low weights to frequent terms

\[ tfidf(t, d, D) = tf(t, d) \times idf(t, D) \]
**WEIGHTING SCHEMES: EXAMPLE**

**Example:** number of documents in corpus (5)

- **Term frequency**
  - high weights: English, speak, feel (2 occurrences)

- **Inverse Document Frequency:**
  - high weights: non frequent terms (patient, hardworking, engineering, etc.)
  - \( \text{idf(India, Corpus)} = \log \frac{5}{1} = 0.69 \)

- **Term frequency-Inverse Document Frequency:**
  - \( \text{tf-idf (India)} = \text{TF(India, d2)} \times \text{IDF (India, Corpus)} \)

---

<table>
<thead>
<tr>
<th>Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1 \rightarrow {It is necessary to be determined, patient and hardworking when learning English}</td>
</tr>
<tr>
<td>d2 \rightarrow {I am 22 years old, engineering student from India}</td>
</tr>
<tr>
<td>d3 \rightarrow {my family cannot speak English}</td>
</tr>
<tr>
<td>d4 \rightarrow {I feel bad to speak with my American girlfriend because I think I could be wrong}</td>
</tr>
<tr>
<td>d5 \rightarrow {I think that we fail in learning and conversing efficiently because we are just feel shy to speak to someone}</td>
</tr>
</tbody>
</table>
**Similarity/Distance Metrics**

- **Cosine**
  - measure of similarity between two vectors of attributes $A$ and $B$
  
  \[
  \text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}
  \]

- **Jaccard Coefficient**
  - measure of similarity between finite sample sets
  - defined as the size of the intersection divided by the size of the union of the sample sets
  
  \[
  J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.
  \]
ENERGY AND VARIANCE

- Energy method
  - retaining enough singular values to make up 90% of the energy in initial matrix
  - sum of the squares of the retained singular values should be at least 90% of the sum of the squares of all the singular values

- Variance method
  - measuring probability distribution
  - plotting singular values in a scree plot
  - retaining dimensionality associated with the knee of the curve
**P-value**

- **Definition**
  - function of observed sample results used for testing a statistical hypothesis
  - significance level of the test, traditionally 5% or 1%

- **Example of a p-value computation**
  - $p$-value is the area under the curve past the observed data point
Regression Analysis

Regression analysis

- Multiple regression analysis is a powerful technique used for predicting the unknown value of a variable from the known value of two or more variables—also called the predictors.
- More than one independent variable, regression line cannot be visualized in the two-dimensional space.
- Predict the value of Y for given values of $X_1, X_2, ..., X_k$

\[ Y = a + b_1X_1 + b_2X_2 + b_3X_3 \ldots b_nX_n \]
Regression Analysis

Regression analysis
- statistical relationship between one or more independent variables and a dependent variable
- mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant
- more than one independent variable, regression line cannot be visualized in the two dimensional space

Coefficient interpretation
- testing p-value for null hypothesis
- low p-value (< 0.05): reject null hypothesis
- larger (insignificant) p-value: no association between dependent and independent variables