Clase 1: Aspectos Generales (Parte B)

Enfoques Clásicos y Neuronales a la Minería de Texto

Marcelo Errecalde[1,2]

[1]Universidad Nacional de San Luis, Argentina

[2]Universidad Nacional de la Patagonia Austral, Argentina

Resumen

1. **Representación de Documentos**
   - Atributos Estáticos
   - Bolsa de Palabras
   - Enfoques Distribucionales
     - Aprendizaje de representaciones
2. **Nuestros trabajos recientes en el área**
   - Detección anticipada de riesgos: casos de depresión en la Web
3. **Líneas de trabajo presentes y futuras**

Representación de Documentos

Representaciones con características estáticas

Se eligen antes del procesamiento de los documentos.

Representación de Documentos

Representaciones con características estáticas

Se eligen antes del procesamiento de los documentos.

Representaciones con características dinámicas

Surgen como parte del procesamiento de los documentos.

Representación de Documentos

Representaciones con características estáticas

Se eligen antes del procesamiento de los documentos.

Representaciones con características dinámicas

Surgen como parte del procesamiento de los documentos.

Representaciones aprendidas

Extienden el uso del aprendizaje automático, también a la representación de los documentos

Ejemplo: Textos arbitrarios - características estáticas

Documentos

1. "pintaron el banco de la plaza"
2. "te paso el programa, ejecútalo paso por paso"
3. "sentado en el banco, miraba si el banco abría"

Ejemplo: Textos arbitrarios - características estáticas

Documentos

1. "pintaron el banco de la plaza"
2. "te paso el programa, ejecútalo paso por paso"
3. "sentado en el banco, miraba si el banco abría"

Número de palabras (NP), longitud de palabra más larga (LPL), longitud promedio de palabras (LPP), verbos en pasado (VP)

ID	NP	LPL	LPP	VP
d1	6	8	4	1
d2	8	9	4,5	0
d3	9	7	4	1

Características estáticas: ejemplo en Wikipedia

Information Processing and Management 54 (2018) 1169–1181

ELSEVIER

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Quality flaw prediction in Spanish Wikipedia: A case of study with verifiability flaws

Edgardo Ferretti[a,b], Leticia Cagnina[*,a,b,c], Viviana Paiz[a], Sebastián Delle Donne[a], Rodrigo Zacagnini[a], Marcelo Errecalde[a,b]

[a] Departamento de Informática, Universidad Nacional de San Luis (UNSL), Ejército de los Andes 950, San Luis, Argentina
[b] Laboratorio de Investigación y, Desarrollo en Inteligencia Computacional (UNSL), Argentina
[c] Consejo Nacional de Investigaciones, Científicas y Técnicas (CONICET), Argentina

Características estáticas: ejemplo en Wikipedia

Table 1
Features that comprise the document model.

Feature	Description
Content-based	
Character count	Number of characters in the text (no spaces).
Word count	Number of words in the plain text.
Sentence count	Number of sentences in the plain text.
Word length	Average word length in characters.
Sentence length	Average sentence length in words.
Paragraph count	Number of paragraphs.
Paragraph length	Average paragraph length in sentences.
Longest word length	Length in characters of the longest word.
Longest sentence length	Number of words in the longest sentence.
Shortest sentence length	Number of words in the shortest sentence.
Long sentence rate	Percentage of long sentences. A long sentence is defined as containing at least 30 words.
Short sentence rate	Percentage of short sentences. A short sentence is defined as containing at most 15 words.
Longest subsection length	Length in words of the longest subsection.
Shortest subsection length	Length in words of the shortest subsection.
Subsections length	Total number of words in the article's subsections.
Average subsection length	Average number of words per subsection.
Longest subsubsection	Length in words of the longest subsubsection.
Shortest subsubsection	Length in words of the shortest subsubsection.
Subsubsections length	Total number of words in the article's subsubsections.
Average subsubsections	Average number of words per subsubsection.
Structure-based	
Section count	Number of sections.
Subsection count	Number of subsections.
Subsubsection count	Number of subsubsections.
Heading count	Number of sections, subsections and subsubsections.
Section nesting	Average number of subsections per section.
Subsection nesting	Average number of subsubsections per subsection.
Reference Sections Count	Number of reference sections, e.g. "References", "Footnotes", "Sources", "Bibliography".
Mandatory Sections Count	Number of mandatory sections, e.g. "See also".
Related page count	Number of related pages, e.g. "Further reading", "See also", etc.
Lead length	Number of words in the lead section (text before the first heading).
Lead rate	Percentage of words in the lead section.
Image count	Number of images.
Image rate	Ratio of image count to section count.
Link count	Every occurrence of a link (introduced with two open square brackets) in the unfiltered text.
Link rate	Percentage of links.
Table count	Number of tables.
Reference count	Number of all references using the `<ref>... </ref>` syntax.
Reference section rate	Ratio of reference count to the accumulated section, subsection and subsubsection count.
Reference word rate	Ratio of reference count to word count.
Unique reference count	Number of unique references using the `<ref>... </ref>` syntax.
Reference ratio	Ratio between the reference word rate of the article and the maximum reference word rate found in the dataset.
Templates-count	Number of (different) Wikipedia templates.

Ejemplo de características dinámicas: Textos arbitrarios (Representación BOW)

- Objetos/instancias: información de cada texto.

Ejemplo de características dinámicas: Textos arbitrarios (Representación BOW)

- Objetos/instancias: información de cada texto.
- Atributos: las distintas palabras que aparecen en los documentos.

Ejemplo de características dinámicas: Textos arbitrarios (Representación BOW)

- Objetos/instancias: información de cada texto.
- Atributos: las distintas palabras que aparecen en los documentos.
- Valores de los atributos: de tipo numéricos o booleanos.

Ejemplo de características dinámicas: Textos arbitrarios (Representación BOW)

- Objetos/instancias: información de cada texto.
- Atributos: las distintas palabras que aparecen en los documentos.
- Valores de los atributos: de tipo numéricos o booleanos.

texto1.txt

```
"pintaron el banco de la plaza"
```

texto2.txt

```
"te paso el programa, ejecútalo paso por paso"
```

texto3.txt

```
"sentado en el banco, miraba si el banco
abría"
```

Ejemplo de características dinámicas: Textos arbitrarios (Representación BOW)

Documentos

1. "pintaron el banco de la plaza"
2. "te paso el programa, ejecútalo paso por paso"
3. "sentado en el banco, miraba si el banco abría"

Ejemplo de características dinámicas: Textos arbitrarios (Representación BOW)

Documentos

1. "pintaron el banco de la plaza"
2. "te paso el programa, ejecútalo paso por paso"
3. "sentado en el banco, miraba si el banco abría"

Pesos binarios

ID	abría	banco	de	ejecútalo	el	en	la	miraba	paso	pintaron	plaza	por	programa	sentado	si	te
d1	0	1	1	0	1	0	1	0	0	1	1	0	0	0	0	0
d2	0	0	0	1	1	0	0	0	1	0	0	1	1	0	0	1
d3	1	1	0	0	1	1	0	1	0	0	0	0	0	1	1	0

Ejemplo de características dinámicas: Textos arbitrarios (Representación BOW)

Documentos

1. "pintaron el banco de la plaza"
2. "te paso el programa, ejecútalo paso por paso"
3. "sentado en el banco, miraba si el banco abría"

Pesos *TF* (Frecuencia del término)

ID	abría	banco	de	ejecútalo	el	en	la	miraba	paso	pintaron	plaza	por	programa	sentado	si	te
d1	0	1	1	0	1	0	1	0	0	1	1	0	0	0	0	0
d2	0	0	0	1	1	0	0	0	3	0	0	1	1	0	0	1
d3	1	2	0	0	2	1	0	1	0	0	0	0	0	1	1	0

Ejemplo de características dinámicas: Textos arbitrarios - *n*-gramas (palabras)

Documentos

1. "pintaron el banco de la plaza"
2. "te paso el programa, ejecútalo paso por paso"
3. "sentado en el banco, miraba si el banco abría"

Ejemplo de características dinámicas: Textos arbitrarios - *n*-gramas (palabras)

Documentos

1. "pintaron el banco de la plaza"
2. "te paso el programa, ejecútalo paso por paso"
3. "sentado en el banco, miraba si el banco abría"

bi-gramas de palabras - Pesos binarios

ID	banco abría	banco de	banco miraba	de la	ejecútalo paso	el banco	el programa	en el	la plaza	...
d1	0	1	0	1	0	1	0	0	1	...
d2	0	0	0	0	1	0	1	0	0	...
d3	1	0	1	0	0	2	0	1	0	...

Ejemplo de características dinámicas: Textos arbitrarios - *n*-gramas (caracteres)

Documentos

1. "pintaron el banco de la plaza"
2. "te paso el programa, ejecútalo paso por paso"
3. "sentado en el banco, miraba si el banco abría"

Ejemplo de características dinámicas: Textos arbitrarios - *n*-gramas (caracteres)

Documentos

1. "pintaron el banco de la plaza"
2. "te paso el programa, ejecútalo paso por paso"
3. "sentado en el banco, miraba si el banco abría"

tri-gramas de caracteres - Pesos binarios

ID	_ab	_ba	_de	_ej	_el	...	aso	aza	ba_	ban	brí	...	tad	tal	tar	te_	úta
d1	0	1	1	0	1	...	0	1	0	1	0	...	0	0	1	0	0
d2	0	0	0	1	1	...	1	0	0	0	0	...	0	1	0	1	1
d3	1	1	0	0	1	...	0	0	1	1	1	...	1	0	0	0	0

Consideraciones finales sobre BoW

Ideas subyacentes

- No se captura ningún tipo de información sobre el orden en que aparecen los términos/palabras

Consideraciones finales sobre BoW

Ideas subyacentes

- No se captura ningún tipo de información sobre el orden en que aparecen los términos/palabras
- BoW sólo mira a la forma superficial de las palabras, ignorando toda información semántica de las mismas

Consideraciones finales sobre BoW

Ideas subyacentes

- No se captura ningún tipo de información sobre el orden en que aparecen los términos/palabras
- BoW sólo mira a la forma superficial de las palabras, ignorando toda información semántica de las mismas

Ventajas

- Simplicidad.
- Eficiencia.

Desventajas

- BoW tiende a producir representaciones muy dispersas (“sparse”)
- Problemas “semánticos” con la polisemia y la sinonimia

Polisemia (y homonimia)

BoW y la polisemia

La polisemia introduce ruido en la representación BoW

Sinonimia

BoW y la sinonimia

La sinonimia divide la evidencia en la representación BoW

Representación distribucional de términos

- Las ideas surgen de las dificultades que se presentan para definir el significado de las palabras.

Representación distribucional de términos

- Las ideas surgen de las dificultades que se presentan para definir el significado de las palabras.
  - Circularidad de las definiciones de diccionario.

Representación distribucional de términos

- Las ideas surgen de las dificultades que se presentan para definir el significado de las palabras.
  - Circularidad de las definiciones de diccionario.
  - Dificultad para capturar otros tipos de relaciones entre palabras (más alla de las relaciones semanticas clásicas de WN) (asociación de palabras, campos semánticos, significados afectivos/connotaciones).

Representación distribucional de términos

- Las ideas surgen de las dificultades que se presentan para definir el significado de las palabras.
  - Circularidad de las definiciones de diccionario.
  - Dificultad para capturar otros tipos de relaciones entre palabras (más alla de las relaciones semanticas clásicas de WN) (asociación de palabras, campos semánticos, significados afectivos/connotaciones).
- Identificadas por filósofos como Ludwig Wittgenstein:

  ... the meaning of a word is its use in the language

Representación distribucional de términos

- Las ideas surgen de las dificultades que se presentan para definir el significado de las palabras.
  - Circularidad de las definiciones de diccionario.
  - Dificultad para capturar otros tipos de relaciones entre palabras (más alla de las relaciones semanticas clásicas de WN) (asociación de palabras, campos semánticos, significados afectivos/connotaciones).
- Identificadas por filósofos como Ludwig Wittgenstein:

  ... the meaning of a word is its use in the language

- ... y lingüistas como John R. Firth:

  You shall know a word by the company it keeps!

Representación distribucional de términos

- Las ideas surgen de las dificultades que se presentan para definir el significado de las palabras.
  - Circularidad de las definiciones de diccionario.
  - Dificultad para capturar otros tipos de relaciones entre palabras (más alla de las relaciones semanticas clásicas de WN) (asociación de palabras, campos semánticos, significados afectivos/connotaciones).
- Identificadas por filósofos como Ludwig Wittgenstein:

  ... the meaning of a word is its use in the language

- ... y lingüistas como John R. Firth:

  You shall know a word by the company it keeps!

Representación distribucional de términos

- Propuesto para abordar las deficiencias del modelo BoW
- El foco se pone en las palabras y los contextos en que éstas ocurren

Representación distribucional de términos

- Se cambia el foco en medir similitud de palabras
- Hipótesis distribucional: palabras que ocurren en contextos similares tienden a tener significados similares.

Representación distribucional de términos

- Cada palabra es representada por un vector
- Cada elemento del vector se deriva de la ocurrencia de la palabra en distintos contextos: otras palabras, frases, sentencias, párrafos, capítulos, documentos, etc.

Representación distribucional de términos

- Cada contexto puede ser visualizado como un concepto
- Cada fila de la matriz palabra-concepto puede ser visualizada entonces como una bolsa de conceptos (BOC)

Algunas representaciones distribucionales conocidas

Existen distintas variantes de las representaciones distribucionales.

Algunas representaciones distribucionales conocidas

Existen distintas variantes de las representaciones distribucionales.

- Document occurrence representation (DOR)
- Term co-occurrence representation (TCOR)
- Concise semantic analysis (CSA)

Veremos ahora, el caso en que estas representaciones de los términos son aprendidas

Características aprendidas

1. Idea: extender el aprendizaje automático, usualmente usado para generar el modelo de clasificación, a la representación de los documentos.

Características aprendidas

1. Idea: extender el aprendizaje automático, usualmente usado para generar el modelo de clasificación, a la representación de los documentos.
2. Componente principal en el aprendizaje de representaciones para el PLN: word embedding (WE) (incrustación de palabra)

Características aprendidas

1. Idea: extender el aprendizaje automático, usualmente usado para generar el modelo de clasificación, a la representación de los documentos.
2. Componente principal en el aprendizaje de representaciones para el PLN: word embedding (WE) (incrustación de palabra)
3. Los WEs son representaciones distribuidas de palabras basadas en vectores densos, de longitud fija construidas mediante estadísticas de co-ocurrencia de palabras según la hipótesis distribucional

Características aprendidas (II)

1. Los WEs pueden ser derivados mediante enfoques basados en conteo:
   - LSA
   - HAL
   - COALS
   - ... y Glove (entre otros)
2. Sin embargo, se han popularizado últimamente los enfoques basados en predicción utilizando métodos de aprendizaje neuronales: word2vec, fastext, Elmo y Bert
3. Los WEs, además de capturar relaciones sintácticas y semánticas muy interesantes de las palabras, soportan el funcionamiento de modelos neuronales más generales para documentos como las arquitecturas LSTM y CNN.

Word2vec

word2vec [Mikolov et. al]

Disponible en:

https://code.google.com/archive/p/word2vec/

Word2vec

word2vec [Mikolov et. al]

Disponible en:

https://code.google.com/archive/p/word2vec/

Word2vec

- El método más popular de embedding en la actualidad.

Word2vec

- El método más popular de embedding en la actualidad.
- A diferencia de los enfoques distribucionales, la idea es **predecir** más que **contar**

Word2vec

- El método más popular de embedding en la actualidad.
- A diferencia de los enfoques distribucionales, la idea es **predecir** más que **contar**
- Se aprende un predictor (clasificador), y como efecto colateral se obtienen los “embeddings” (vectores) que representan las palabras.

Word2vec

- El método más popular de embedding en la actualidad.
- A diferencia de los enfoques distribucionales, la idea es **predecir** más que **contar**
- Se aprende un predictor (clasificador), y como efecto colateral se obtienen los “embeddings” (vectores) que representan las palabras.
- Enfoque bastante eficiente para aprender

Enfoques en Word2vec

Evaluación de embeddings

Evaluación extrínseca sobre *otras tareas*

Agregarlos como “features” en otras tareas de NLP (análisis de sentimiento, perfilado de autor, etc) y ver si esto mejora el desempeño sobre algun otro modelo

Evaluación intrínseca (*similitud* entre palabras)

Correlación con ratings de similitud asignados por humanos

- WordSim-353 (Finkelstein et al., 2002)
- SimLex-999 (Hill et al., 2015)
- Stanford Contextual Word Similarity (SCWS) dataset (Huang et al., 2012)
- TOEFL dataset: `Levied is closest in meaning to: imposed, believed, requested, correlated`

Visualización de embeddings

Proyección bi-dimensional (t-SNE) - dimensión original 60 ([Li et al. (2015)])

Visualización de embeddings (II)

Propiedades (semánticas) de los embeddings

Tamaños de la ventana de contexto (C):

- Cortos: las palabras más similares a la palabra objetivo son semánticamente similares y con el mismo POS.
- Largos: las palabras más similares a la palabra objetivo están relacionadas al tópico pero no son similares.

Propiedades (semánticas) de los embeddings

Habilidad para capturar significados relacionales!!!!:

- *vector*('king') - *vector*('man') + *vector*('woman') ≈ *vector*('queen')
- *vector*('Paris') - *vector*('France') + *vector*('Italy') ≈ *vector*('Rome')

Propiedades (semánticas) de los embeddings (II)

Cambios en el significado (histórico) de las palabras:

Embeddings y arquitectuas recurrentes

RNN/LSTM en clasificación

RNN/LSTM en clasificación

RNN/LSTM en traducción

Estado actual: la “carrera” de los “transformers”

- Junio de 2017. Arquitectura Transformer Google AI
- Mayo de 2018. ULMFIT fast.ai
- Noviembre de 2018. BERT Google AI
- Enero de 2019. Transformer-XL Google AI
- Enero de 2019. Modelos Neur. de StanfordNLP
- Febrero de 2019. GPT-2 OpenAI
- Junio de 2019. XLNet CMUAI
- Julio de 2019. PyTorch-Transformers
- Julio de 2019. ERNIE Baidu 百度
- Agosto de 2019. RoBERTa facebook research
- Agosto de 2019. Spacy-PyTorch-Transformers spaCy
- Agosto de 2019. XLM/mBERT facebook research

Nuestros trabajos recientes en Minería de Textos

- Agrupamiento de textos cortos
- Calidad de información en la Web (esp. en Wikipedia)
- Perfil del autor en medios sociales (edad, género, rasgos de personalidad, etc)
- Clasificación temprana de textos/Detección temprana de riesgos

(Agrupamiento de textos cortos - técnicas bio-inspiradas

Técnicas basadas en hormigas

Técnicas basadas en PSO

Publicaciones recientes

Calidad de Información en Wikipedia

Artículos Destacados

Detección de Fallas

Visualización

Métricas de Calidad

Detección de fallas en Wikipedia en español

Information Processing and Management 54 (2018) 1169–1181

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Quality flaw prediction in Spanish Wikipedia: A case of study with verifiability flaws

Edgardo Ferretti[a,b], Leticia Cagnina[*,a,b,c], Viviana Paiz[a], Sebastián Delle Donne[a], Rodrigo Zacagnini[a], Marcelo Errecalde[a,b]

[a] Departamento de Informática, Universidad Nacional de San Luis (UNSL), Ejército de los Andes 950, San Luis, Argentina
[b] Laboratorio de Investigación y, Desarrollo en Inteligencia Computacional (UNSL), Argentina
[c] Consejo Nacional de Investigaciones, Científicas y Técnicas (CONICET), Argentina

Detección de fallas en Wikipedia en español: features

Table 1
Features that comprise the document model.

Feature	Description
Content-based	
Character count	Number of characters in the text (no spaces).
Word count	Number of words in the plain text.
Sentence count	Number of sentences in the plain text.
Word length	Average word length in characters.
Sentence length	Average sentence length in words.
Paragraph count	Number of paragraphs.
Paragraph length	Average paragraph length in sentences.
Longest word length	Length in characters of the longest word.
Longest sentence length	Number of words in the longest sentence.
Shortest sentence length	Number of words in the shortest sentence.
Long sentence rate	Percentage of long sentences. A long sentence is defined as containing at least 30 words.
Short sentence rate	Percentage of short sentences. A short sentence is defined as containing at most 15 words.
Longest subsection length	Length in words of the longest subsection.
Shortest subsection length	Length in words of the shortest subsection.
Subsections length	Total number of words in the article's subsections.
Average subsection length	Average number of words per subsection.
Longest subsubsection	Length in words of the longest subsubsection.
Shortest subsubsection	Length in words of the shortest subsubsection.
Subsubsections length	Total number of words in the article's subsubsections.
Average subsubsections	Average number of words per subsubsection.
Structure-based	
Section count	Number of sections.
Subsection count	Number of subsections.
Subsubsection count	Number of subsubsections.
Heading count	Number of sections, subsections and subsubsections.
Section nesting	Average number of subsections per section.
Subsection nesting	Average number of subsubsections per subsection.
Reference Sections Count	Number of reference sections, e.g. "References", "Footnotes", "Sources", "Bibliography".
Mandatory Sections Count	Number of mandatory sections, e.g. "See also".
Related page count	Number of related pages, e.g. "Further reading", "See also", etc.
Lead length	Number of words in the lead section (text before the first heading).
Lead rate	Percentage of words in the lead section.
Image count	Number of images.
Image rate	Ratio of image count to section count.
Link count	Every occurrence of a link (introduced with two open square brackets) in the unfiltered text.
Link rate	Percentage of links.
Table count	Number of tables.
Reference count	Number of all references using the `<ref>... </ref>` syntax.
Reference section rate	Ratio of reference count to the accumulated section, subsection and subsubsection count.
Reference word rate	Ratio of reference count to word count.
Unique reference count	Number of unique references using the `<ref>... </ref>` syntax.
Reference ratio	Ratio between the reference word rate of the article and the maximum reference word rate found in the dataset.
Templates-count	Number of (different) Wikipedia templates.

Detección de fallas en Wikipedia: PU-Learning

Figure 2. Non-iterative proposed approach for the two-step strategy to PU Learning

"On the Use of PU Learning for Quality Flaw Prediction in Wikipedia". Edgardo Ferretti, Donato Hernández, Rafael Guzman, Manuel, Montes-Y-Gómez, Marcelo Errecalde and Paolo Rosso. CLEF2012 - Labs. Track: PAN - Quality Flaw Prediction in Wikipedia. 2012

"On the Use of Reliable-Negatives Selection Strategies in the PU Learning Approach for Quality Flaws Prediction in Wikipedia". Edgardo Ferretti, Marcelo Errecalde, Maik Anderka and Benno Stein. In IEEE Proceedings of DEXA-2014 Publisher: IEEE, Munich, Germany, 2014.

Métricas de Calidad en Wikipedia:: Información factual

"Measuring the Quality of Web Content using Factual Information". E. Lex, M. Voelske, M. Errecalde, E. Ferretti, L. Cagnina, Ch. Horn, M. Granitzer, B. Stein. Proc. of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality, 2012. ACM, New York, NY, USA.
"On the Feasibility of External Factual Support as Wikipedia's Quality Metric", C. G. Velázquez, L. C. Cagnina, M. L. Errecalde. Revista del Procesamiento del Lenguaje Natural (SEPLN), Vol. 58, pp. 93-100. 2017.

Análisis de Calidad Interactivo: Wikilizer

Fig. 1. *Quality Agent* UI. (A) QM Panel: contains built-in and custom QMs. (B) Attributes Panel: displays all available attributes to create QMs. (C) Ranking View: ranks articles in the current dataset by the selected QM(s). (D) Equation Composer: area where experts can create QMs through mathematical combinations of attributes and/or other QMs. (E) Article Panel: shows quality-based evolution of the selected article (top) and its mobile version (below).

"WikiLyzer: Interactive Information Quality Assessment in Wikipedia". Cecilia di Sciascio, David Strohmaier, Marcelo Errecalde,. Proceedings of the 22nd International Conference on Intelligent User Interfaces. Pages 377-388. Limassol, Cyprus - March 13 - 16, 2017. ACM New York, NY, USA, 2017.

Perfil de autor

Tareas: determinar

1. edad
2. género
3. personalidad
4. orientación política
5. ...

Perfil de autor

"Cross Domain Author Profiling Task in Spanish Language: An Experimental Study". María José Garciarena Ucelay, María Paula Villegas, Leticia C. Cagnina and Marcelo L. Errecalde. Journal of Computer Science and Technology (JCST). Vol. 15. Nro. 2, pp. 122-128. Iberoamerican Science & Technology Education Consortium. México/Argentina. ISSN: 1666-6038, November 2015.

Detección de pedófilos en la Web

- Datos de entrenamiento en www.perverted-justice.com
- Competencias recientes pan.webis.de/clef12/pan12-web

Ejemplo:

Detección de pedófilos en la Web

- Enfoque: utilizar un perfil de n-gramas de caracteres con información de las etapas y una adecuada función de similitud para detectar los cambios de escritura.

Detección de pedófilos en la Web

- Enfoque: utilizar un perfil de n-gramas de caracteres con información de las etapas y una adecuada función de similitud para detectar los cambios de escritura.

"Character n-grams profiles for predatory lines detection". Leticia C. Cagnina, Marcelo L. Errecalde and Efstathios Stamatatos. Proceedings del Workshop de Procesamiento Automatizado de Textos y Corpus 2016 (WOPATEC-2016),

Detección de pedófilos en la Web

Detección de depredadores sexuales en chats

1) **Elaboración del perfil de la etapa (perfil del depredador):**

Grooming

Líneas catalogadas extraídas de conversaciones reales de pedófilos.

Hi how are you?
Have you been? Have you any plan?
How old are you? Are you alone?
Give me your address.
I want see you tonight
You are perfect for me. I like you.
Do you want to see me tomorrow?
Give me you opinion about this picture.
Where do you think is a good place?

3gramas de caracteres

Hi_
i_h
ho_
how
ow_
w_a
ar_
are
re_
….

2) **Elaboración del perfil de la conversación**

Procesar la conversación utilizando "ventanas móviles".

Elaborar el perfil de la ventana.

{Adamou217}	whats ur screen name..mines adamou217..on either
{superboy_93	superboy93
Adamou217	who's tim
DFW Superb	my best firend
Adamou217	ah the bi one
Adamou217	u boys ever play together..lol
DFW Super	yea we have ;)
Adamou21	
Adamou217	when do i get a chance..lol
Adamou217	is she home tonight
DFW Superboy 93	so far
DFW Superboy 93	she has to work tomorow
Adamou217	ah well maybe she'll go to be early..:)

Detección de pedófilos en la Web

Detección de depredadores sexuales en chats

3) Buscar similitudes entre el perfil de la ventana y el perfil del depredador

Clasificación temprana/anticipada

- Entrenar con información secuencial completa.
- Luego clasificar, tan pronto como sea possible

Clasificación temprana/anticipada

- Entrenar con información secuencial completa.
- Luego clasificar, tan pronto como sea possible

"Early text classification: a Naive solution". Hugo Jair Escalante, Manuel Montes y Gómez, Luis Villaseñor-Pineda and Marcelo Luis Errecalde. Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2016), pp. 91-99, ISBN 978-1-941643-82-2, NAACL-HLT 2016, Association for Computational Linguistics, San Diego, California, June 2016. URL: www.aclweb.org/anthology/W16-0416

El AAT en las Ciencias Sociales

- Enfoque tradicional: estudios controlados, diseñados minuciosamente y a baja escala (número pequeño de voluntarios).

El AAT en las Ciencias Sociales

- Enfoque tradicional: estudios controlados, diseñados minuciosamente y a baja escala (número pequeño de voluntarios).
- Por otro lado, en la era de la información, Internet hace disponible grandes volúmenes de contenido textual de:

El AAT en las Ciencias Sociales

- Enfoque tradicional: estudios controlados, diseñados minuciosamente y a baja escala (número pequeño de voluntarios).
- Por otro lado, en la era de la información, Internet hace disponible grandes volúmenes de contenido textual de:
  - Medios Sociales (como Facebook o Twitter): usados por más de 1.000.000.000 de personas.
  - Sitios de noticias
  - Contenidos digitales on-line
  - Plataformas de e-learning

Idea

Usar técnicas de análisis automáticos de textos (AAT) (a.k.a. análisis automático de contenido) para entender a las personas a una escala sin precedentes.

Varios cambios metodológicos:

- análisis masivo de contenidos dirigido por los datos
- teorías a priori (vocabularios controlados) versus asociaciones inesperadas (vocabularios abiertos)

Depresión: un flagelo de nuestros días

La depresión en la salud pública

- En 2015, se estimó un 4,4 % de la población mundial con depresión (más de 332 millones de personas).

Depresión: un flagelo de nuestros días

La depresión en la salud pública

- En 2015, se estimó un 4,4 % de la población mundial con depresión (más de 332 millones de personas).
- Más allá de los efectos perjudiciales que tiene la depresión, ésta puede llevar al suicidio

Depresión: un flagelo de nuestros días

La depresión en la salud pública

- En 2015, se estimó un 4,4 % de la población mundial con depresión (más de 332 millones de personas).
- Más allá de los efectos perjudiciales que tiene la depresión, ésta puede llevar al suicidio
- Más de 800.000 muertes por suicidio por año y segunda causa de muerte en el rango de 15 a 29 años.

Depresión: un flagelo de nuestros días

La depresión en la salud pública

- En 2015, se estimó un 4,4 % de la población mundial con depresión (más de 332 millones de personas).
- Más allá de los efectos perjudiciales que tiene la depresión, ésta puede llevar al suicidio
- Más de 800.000 muertes por suicidio por año y segunda causa de muerte en el rango de 15 a 29 años.

Depresión: un flagelo de nuestros días

La depresión en la salud pública

- En 2015, se estimó un 4,4 % de la población mundial con depresión (más de 332 millones de personas).
- Más allá de los efectos perjudiciales que tiene la depresión, ésta puede llevar al suicidio
- Más de 800.000 muertes por suicidio por año y segunda causa de muerte en el rango de 15 a 29 años.

En este contexto, la investigación de métodos que asistan en la detección de posibles casos de depresión se vuelve de gran importancia

Algunos enfoques de detección automática de depresión (DAD)

- Stirman y Pennebaker, en 2001, usan Linguistic Inquiry and Word Count (LIWC) para caracterizar depresión a partir del uso del lenguaje natural (patrones interesantes con pronombres)

Algunos enfoques de detección automática de depresión (DAD)

- Stirman y Pennebaker, en 2001, usan Linguistic Inquiry and Word Count (LIWC) para caracterizar depresión a partir del uso del lenguaje natural (patrones interesantes con pronombres)
- De Choudhury y otros, en 2013, analizan tweets de sujetos diagnosticados de depresións, entrenando con SVM. Cuentan con scores de tests de depresión (CES-D y BDI), información de la red de contactos, índices de insomnio, etc.)

Algunos enfoques de detección automática de depresión (DAD)

- Stirman y Pennebaker, en 2001, usan Linguistic Inquiry and Word Count (LIWC) para caracterizar depresión a partir del uso del lenguaje natural (patrones interesantes con pronombres)
- De Choudhury y otros, en 2013, analizan tweets de sujetos diagnosticados de depresións, entrenando con SVM. Cuentan con scores de tests de depresión (CES-D y BDI), información de la red de contactos, índices de insomnio, etc.)

Detección anticipada de riesgos en contexto

Detección anticipada de riesgos (DAR)

Tarea donde los datos son leídos secuencialmente como un flujo de datos y el desafío consiste en detectar casos de riesgo, tan pronto como sea posible.

Detección anticipada de riesgos en contexto

Detección anticipada de riesgos (DAR)

Tarea donde los datos son leídos secuencialmente como un flujo de datos y el desafío consiste en detectar casos de riesgo, tan pronto como sea posible.

Aplicaciones Potenciales

Detección temprana de:

- predadores sexuales
- personas con inclinaciones suicidas
- personas con signos de depresión
- comportamiento agresivo en las redes sociales
- actividades criminales y terroristas
- abandono en cursos académicos

Una tarea internacional dedicada al tema .....

Ediciones de eRisk

eRisk 2017 (https://early.irlab.org/2017/index.html)

- Task: Early Detection of Depression

Ediciones de eRisk

eRisk 2017 (https://early.irlab.org/2017/index.html)

- Task: Early Detection of Depression

eRisk 2018 (https://early.irlab.org/2018/index.html)

- Task 1: Early Detection of Signs of Depression
- Task 2: Early Detection of Signs of Anorexia

Ediciones de eRisk

eRisk 2017 (https://early.irlab.org/2017/index.html)

- Task: Early Detection of Depression

eRisk 2018 (https://early.irlab.org/2018/index.html)

- Task 1: Early Detection of Signs of Depression
- Task 2: Early Detection of Signs of Anorexia

eRisk 2019 (https://early.irlab.org/)

- Task 1: Early Detection of Signs of Anorexia
- Task 2: Early Detection of Signs of Self-harm
- Task 3: Measuring the severity of the signs of depression

La tarea ERISK 2017

Primera edición de la tarea

Erisk 2017 - CLEF 2017 Workshop

Conference and Labs of the Evaluation Forum

La tarea ERISK 2017

Primera edición de la tarea

Erisk 2017 - CLEF 2017 Workshop

Conference and Labs of the Evaluation Forum

2) Nuestra participación se realiza con un método diseñado específicamente para este tipo de tarea (TVT)

La tarea ERISK 2017

Primera edición de la tarea

Erisk 2017 - CLEF 2017 Workshop

Conference and Labs of the Evaluation Forum

2) Nuestra participación se realiza con un método diseñado específicamente para este tipo de tarea (TVT)

3) Se obtiene el mejor valor en la medida $ERDE_{50}$

Conjuntos de Datos

Conj. de Entrenamiento

486 usuarios

- 83 positivos (deprimidos)
- 403 negativos (no-dep.)

Conjunto de Prueba

401 usuarios

- 52 positivos (deprimidos)
- 349 negativos (no-dep.)

Conjuntos de Datos

Conj. de Entrenamiento

486 usuarios

- 83 positivos (deprimidos)
- 403 negativos (no-dep.)

Conjunto de Prueba

401 usuarios

- 52 positivos (deprimidos)
- 349 negativos (no-dep.)

Divididos en 10 bloques (ordenados cronológicamente):

Conjunto de prueba, provisto bloque por bloque ...

Conjunto de prueba, provisto bloque por bloque ...

Conjunto de prueba, provisto bloque por bloque ...

Conjunto de prueba, provisto bloque por bloque ...

La tarea eRisk 2018

Enfoques utilizados

- FTVT
- Sequential-Incremental Classification (SIC)

La tarea eRisk 2018

Enfoques utilizados

- FTVT
- Sequential-Incremental Classification (SIC)

Resultados obtenidos

- Mejores valores de $ERDE_5$ tanto en anorexia como depresión
- Valor más alto de precisión en anorexia (0.91)

Nuestro enfoque actual para DAR: SS3

Explicado en ...

Expert Systems With Applications 133 (2019) 182–197

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A text classification framework for simple and effective early depression detection over social media streams

Sergio G. Burdisso[a,b,*], Marcelo Errecalde[a], Manuel Montes-y-Gómez[c]

[a] Universidad Nacional de San Luis (UNSL), Ejército de Los Andes 950, San Luis, San Lius C.P. 5700, Argentina
[b] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina
[c] Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Luis Enrique Erro No. 1, Sta. Ma. Tonantzintla, Puebla C.P. 72840, Mexico

Método flexible para clasificación de casos de riesgo

Reconocimiento de “gránulo fino” de palabras relevantes

Tamaño por valor global (SS3)

Tamaño por frecuencia común

Interpretación gráfica de resultados (caja blanca)

...

Writing 54 ▶ I'm going to agree with everyone else and say you definitely need a lawyer. Get in touch with the[...]

Writing 55 ▶ You don't mention what the fertility issue is (and you don't have to) but his feelings may stem fr[...]

...

Writing 59 ▶ Thankfully I was able to realize that I was in a bad place and get help. My sister has been awesom[...]

Writing 60 ▶ I have been seeing a therapist which I think is helping a little. Fact is, I was feeling really depressed[...]

Writing 61 ▶ My Wife Wants a Divorce . This will be long, sorry in advance. My wife told me shortly after the[...]

Writing 62 ▶ the Earth Arena coming up I have: Zelnite x2 Dilma Ophelia For my last spot, should I use Miku [...]

...

(a) Subject 9579's history - writing level

I have been seeing a therapist which I think is helping a little. Fact is, I was feeling really depressed and wanting to kill myself. I spent basically all of Feb in the hospital[...]

(b) Writing 60 - sentence level

I have been seeing a therapist which I think is helping a little. Fact is, I was feeling really depressed and wanting to kill myself. I spent basically all of Feb in the hospital[...]

(c) Writing 60 - word level

SS3 en vivo (`http://tworld.io/ss3/`)

La tarea eRisk 2019

Enfoques utilizados

- Clasificador de texto SS3
- Estimador de la severidad de los síntomas de depresión (basado en SS3)

La tarea eRisk 2019

Enfoques utilizados

- Clasificador de texto SS3
- Estimador de la severidad de los síntomas de depresión (basado en SS3)

Resultados obtenidos

- SS3 fue el método más rápido
- Obtuvo el mejor *ERDE* y las mejores medidas basadas en ranking en todas las tareas.
- La mejor precisión, F_1 y $F_{latency}$ en la tarea T2
- En la tarea T3, obtuvo los mejores valores de AHR y ACR, y el segundo mejor valor de ADODL y DCHR.

Nueva tarea: completar el cuestionario Beck's Depression Inventory (BDI)

```
Instructions:

This questionnaire consists of 21 groups of statements. Please read each group of statements
carefully, and then pick out the one statement in each group that best describes the way you feel.
If several statements in the group seem to apply equally well, choose the highest
number for that group.

1. Sadness
0. I do not feel sad.
1. I feel sad much of the time.
2. I am sad all the time.
3. I am so sad or unhappy that I can't stand it.

2. Pessimism
0. I am not discouraged about my future.
1. I feel more discouraged about my future than I used to be.
2. I do not expect things to work out for me.
3. I feel my future is hopeless and will only get worse.

3. Past Failure
0. I do not feel like a failure.
1. I have failed more than I should have.
2. As I look back, I see a lot of failures.
3. I feel I am a total failure as a person.

4. Loss of Pleasure
0. I get as much pleasure as I ever did from the things I enjoy.
1. I don't enjoy things as much as I used to.
2. I get very little pleasure from the things I used to enjoy.
3. I can't get any pleasure from the things I used to enjoy.
```

Líneas de trabajo presentes y futuras

Detección anticipada de riesgos

- Pedófilos
- Adictos
- Suicidas
- Multimodal

Líneas de trabajo presentes y futuras

Detección anticipada de riesgos

- Pedófilos
- Adictos
- Suicidas
- Multimodal

Conocimiento y Aprendizaje

- Aprendizaje de embeddings basado en conocimiento
- Semánticas de conceptos y semánticas distribucionales

Líneas de trabajo presentes y futuras

Detección anticipada de riesgos

- Pedófilos
- Adictos
- Suicidas
- Multimodal

Conocimiento y Aprendizaje

- Aprendizaje de embeddings basado en conocimiento
- Semánticas de conceptos y semánticas distribucionales

Aplicaciones PLN-intensivas

- Chatbots de uso comercial
- Extracción de información cerrada
- Análisis de sentimientos basado en aspectos

¿Preguntas?

¿Preguntas?

Muchas gracias por su atención!!!

¿Preguntas?

Muchas gracias por su atención!!!

Información de contacto

Marcelo Errecalde merreca@unsl.edu.ar
merrecalde@gmail.com