Minería de texto y Recuperación de Información

Ideas para proyectos de curso

1. Enriquecer la representación de BoW con expresiones multipalabra para la identificación de lenguaje agresivo

-> Usando el Toolkit mwetoolkit (http://mwetoolkit.sourceforge.net/PHITE.php) definir un conjunto de patrones de expresiones agresivas (vulgares, coloquiales, groseras) para identificar aquellas con mayor frecuencia en el corpus de perfiles MEX-A3T
-> Contrastar la BoW simple contra la BoW + expresiones agresivas para identificar los tuits ofensivos usando un esquema de clasificación supervisada.
-> Posible variante comprobar si dichas expresiones pueden ser útiles para Author Profiling (género, lugar de residencia, edad)

Lecturas sugeridas:

- Kumar, R., Ojha, A.K., Malmasi, S. and Zampieri, M. (2018) Benchmarking Aggression Identification in Social Media. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC). pp. 1-11.
- Waseem, Z., Davidson, T., Warmsley, D. and Weber, I. (2017) Understanding Abuse: A Typology of Abusive Language Detection Subtasks. Proceedings of the Abusive Language Online Workshop.

Foros de evaluación:

- Semeval - Task 5 Multilingual detection of hate speech against immigrants and women in Twitter -https://competitions.codalab.org/competitions/19935
- Semeval - Task 6 Identifying and Categorizing Offensive Language in Social Media https://competitions.codalab.org/competitions/20011

2. Uso de información extraída de las imágenes compartidas en redes sociales para Author Profiling en escenarios multilingües

-> Ya hemos comprobado que es posible usar la información de imágenes para el perfil de usuarios, ahora la idea es comprobar si las imágenes de una colección para un idioma pueden ser útiles para identificar el perfil en otro idioma.
-> Utilizar un sistema ya existente para etiquetar las imágenes con método supervisado (de vocabulario cerrado)
-> Se utilizarán las colecciones del pan 2018 en inglés, español, árabe (https://pan.webis.de/clef18/pan18-web/author-profiling.html)
-> Realizar y contrastar experimentos mono y crosslingues

-> Analizar los resultados en función de la semejanza se los contenidos de las imágenes en esos 3 idiomas (i.e. ¿las mujeres españolas comparten el mismo tipo de imágenes que las árabes?)

Lecturas sugeridas:

- Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter. Francisco Rangel, Paolo Rosso, Manuel Montes-y-Gómez, Martin Potthast, Benno Stein.
- Very deep convolutional networks for large-scale image recognition. Karen Simonyan and Andrew Zisserman.C oRR, abs/1409.1556, 2014

Foros de evaluación:

- PAN 2018 - Author Profiling: https://pan.webis.de/clef18/pan18-web/author-profiling.html

3. Enfoque semi supervisado para detección de agresividad en redes sociales

-> Para abordar esta tarea de forma supervisada es necesario contar con un corpus para la detección de agresividad. Dado el costo de esta tarea, un enfoque es etiquetar automáticamente, usando reglas y diccionarios, un conjunto pequeño de tuits y después aplicar un enfoque self-training

-> Contrastar usando un conjunto inicial etiquetado manualmente.

-> definir los criterios de confianza para incrementar el conjunto de entrenamiento así como el criterio de paro.

-> Un caso interesante es usar word embeddings preentrenados y no preentrenados en un esquema de co-training.

Lecturas sugeridas:

- Kumar, R., Ojha, A.K., Malmasi, S. and Zampieri, M. (2018) Benchmarking Aggression Identification in Social Media. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC). pp. 1-11.
- Waseem, Z., Davidson, T., Warmsley, D. and Weber, I. (2017) Understanding Abuse: A Typology of Abusive Language Detection Subtasks. Proceedings of the Abusive Language Online Workshop.
- Jo, Hwiyeol and Cinarel, Ceyda. Δ-training: Simple Semi-Supervised Text Classification using Pretrained Word Embeddings. (https://arxiv.org/pdf/1901.07651.pdf)

Foros de evaluación:

- Semeval - Task 5 Multilingual detection of hate speech against immigrants and women in Twitter -https://competitions.codalab.org/competitions/19935

- Semeval - Task 6 Identifying and Categorizing Offensive Language in Social Media https://competitions.codalab.org/competitions/20011

4. Diseño de un pesado orientado a la tarea de detección de agresividad

-> En la representación vectorial de los documentos un componente muy importante es el esquema de pesado usado. Los esquemas de pesado más usados son el TF, TF-IDF, TF-IG, con los cuales se intenta medir la capacidad descriptiva y discriminativa de cada término.

-> La idea de este proyecto es proponer algún pesado orientado a la tarea de detección de agresividad, que indique que tan fuerte es uso de esa palabras en textos agresivos. La sugerencia es hacerlo midiendo la cercanía de cada palabra con un diccionario de groserías; esto a través de su representaciones distribuidas (word embeddings)

-> Hacer experimentos combinando nuevo peso con TF, tf-idf, tf-ig.

-> Esta idea también se podría orientar a la tarea de detección de depresión

Lecturas sugeridas:

- Daniel Jurafsky & James H. Martin. Vector Semantics. Book: Speech and Language Processing. Draft of September 23, 2018.
- Waseem, Z., Davidson, T., Warmsley, D. and Weber, I. (2017) Understanding Abuse: A Typology of Abusive Language Detection Subtasks. Proceedings of the Abusive Language Online Workshop.
- Automatic Expansion of Lexicons for Multilingual Misogyny Detection. Simona Frenda, Bilal Ghanem, Estefanía Guzmán-Falcón and Manuel Montes-y-Gómez, Luis Villaseñor-Pineda. 5th Italian Conference on Computational Linguistics (CLiC-it 2018), EVALITA reports, Track on Automatic Misogyny Identification. Turin, Italy, December 10-12, 2018.

Foros de evaluación:

- SemEval - Task 5 Multilingual detection of hate speech against immigrants and women in Twitter -https://competitions.codalab.org/competitions/19935
- SemEval - Task 6 Identifying and Categorizing Offensive Language in Social Media https://competitions.codalab.org/competitions/20011
- Mex-A3T: Authorship and aggressiveness analysis in Mexican Twitter https://mexa3t.wixsite.com/home

5. Explorar sinergia entre AP y detección de depresión

-> Distintos tipos de personas (mujeres y hombres, jóvenes y adultos mayores, norteños y sureños, por ejemplo) suelen tener distintos intereses temáticos, distintas perspectivas sobre mismos asuntos, e incluso a manifestar sus emociones de distinta manera. Así, la idea de este proyecto es explorar el uso de información sobre los usuarios en la detección de comentarios agresivos. Básicamente, la hipótesis es que distintos tipos de personas manifiestan su estado depresivo de distinta manera.

-> La metodología a seguir es: etiquetar mensajes con atributos de sus autores (género, edad, lugar de residencia, por ejemplo) usando algún enfoque de author profiling; aprovechar estas etiquetas como atributos extra o para construir varios clasificadores independientes; entrenar clasificadores supervisados.

Lecturas sugeridas:

- Emphasizing Personal Information for Author Profiling: New Approaches for Term Selection and Weighting. Rosa María Ortega, Adrián Pastor López-Monroy, Anilú Franco, Manuel Montes-y-Gómez. Knowledge-Based Systems. Volume 145, 1 April 2018, Pages 169-181.
- Predicting Depression via Social Media. Munmun De Choudhury Michael Gamon Scott Counts Eric Horvitz.

Foros de evaluación:

- CLEF eRisk 2019: Early risk prediction on the Internet (http://early.irlab.org/)

6. Método de Author Profiling basada en técnicas de enmascaramiento

-> Se tiene evidencia sobre la utilidad de usar técnicas de enmascaramiento para tareas de Atribución de autoría y Detección de engaño. La idea detrás de estas técnicas es seleccionar atributos en función de su frecuencia de uso (en la colección en general, o por las categorías de la tarea), y conservar sólo aquellos relevantes para representar el tipo de información a capturar. Por ejemplo, para atribución de autoría el interés recae en aquellas partículas como las palabras vacías, y se enmascaran las palabras de contenido.

-> El presente proyecto busca evaluar esta técnica sobre la tarea de author profiling, donde los atributos relevantes son una combinación de palabras vacías y de palabras de contenido (tal como se comprobó en la tarea de detección de engaño).

Lecturas sugeridas:

- Stamatatos, Efstathios. "Masking topic-related information to enhance authorship attribution." Journal of the Association for Information Science and Technology 69.3 (2018): 461-473.
- Juan Javier Sánchez-Junquera. Adaptación de Dominio para la Detección Automática de Textos Engañosos. Instituto Nacional de Astrofísica, Óptica y Electrónica. August 29, 2018.

- Juan Javier Sánchez-Junquera, et al. Masking domain-specific information for cross-domain deception detection (in evaluation *Pattern Recognition Letters*)

Foros de evaluación:
- PAN 2018 - Author Profiling: https://pan.webis.de/clef18/pan18-web/author-profiling.html

6. Probar una red convolucional y una recurrente en tareas como detección de depresión y/o agresividad

-> Los modelos basados en redes neuronales son actualmente el estado del arte en varias tareas de clasificación y reconocimiento, aunque poco han sido empleadas en tareas de detección de comportamiento en redes sociales.
-> La idea de este proyecto es aplicar, y comparar, el uso de dos arquitecturas de redes neuronales, una convolucional y otra recurrente, en tareas como detección de depresión y agresividad. la hipótesis es que la convolucional será más apropiada para la detección de depresión, donde el orden de las palabras no es tan relevante, sino su sola presencia, y la recurrente será mejor en la detección de agresividad donde el orden de las palabras es un aspecto importante.

Lecturas sugeridas:
- Convolutional Neural Networks for Authorship Attribution of Short Texts. Prasha Shrestha, Sebastian Sierra, Fabio Gonzalez, Manuel Montes, Paolo Rosso and Thamar Solorio. European Chapter of the Association for Computational Linguistics (EACL 2017). Valencia, 3-7 April 2017.
- Nishant Nikhil, Ramit Pahwa, Mehul Kumar Nirala, and Rohan Khilnani. 2018. Lstms with attention for aggression detection. InProceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC – 1),Santa Fe, USA

Foros de evaluación:
- CLEF eRisk 2019: Early risk prediction on the Internet (http://early.irlab.org/)
- SemEval - Task 5 Multilingual detection of hate speech against immigrants and women in Twitter -https://competitions.codalab.org/competitions/19935
- SemEval - Task 6 Identifying and Categorizing Offensive Language in Social Media https://competitions.codalab.org/competitions/20011
- Mex-A3T: Authorship and aggressiveness analysis in Mexican Twitter https://mexa3t.wixsite.com/home

7. Aprendizaje de representaciones de documentos con programación genética a partir de word embeddings

- Dada una representación distribuida de palabras, aprender a generar la representación de un documento.

- Se sugiere usar programación genética para aprender una función que combine las representaciones de palabras que aparezcan en el documento
- Se espera mejorar el desempeño de representaciones triviales (e.g., promedio) y/o aprendidas (e.g., doc2vec)

Lecturas sugeridas:

- Hugo Jair Escalante, Mauricio A. García-Limón, Alicia Morales-Reyes, Mario Graff, Manuel Montes-y-Gómez, Eduardo F. Morales, José Martínez-Carranza: Term-weighting learning via genetic programming for text classification. Knowl.-Based Syst. 83: 176-189 (2015) https://doi.org/10.1016/j.knosys.2015.03.025
- Distributed Representations of Sentences and Documents Quoc V. Le, Tomas Mikolov, ArXiv, 2014. https://arxiv.org/abs/1405.4053

8. Ordenamiento de textos académicos según su legibilidad

- Aplicar el método de Tanaka, Tezuka y Terada a la comparación de textos académicos a nivel técnico, licenciatura, maestría y doctorado.
- Usar la colección de tesis y propuestas a distintos niveles que ya se ha recopilado.
- Explorar variaciones en características y clasificador.
- Se pretende evaluar textos nuevos para apoyar a estudiantes en el desarrollo de sus escritos.

Lecturas sugeridas:

- TANAKA-ISHII, Kumiko; TEZUKA, Satoshi; TERADA, Hiroshi. Sorting texts by readability. Computational Linguistics, 2010, vol. 36, no 2, p. 203-227.
- González-López, S. , & López-López, A. (2015). Colección de tesis y propuesta de investigación en tics: un recurso para su análisis y estudio. In XIII Congreso Nacional de Investigación Educativa (pp. 1-15).