SOLICITUDE DE APROBACIÓN DE ANTEPROXECTO DO TRABALLO DE FIN DE MÁSTER

Máster universitario en Tecnologías de Análisis de Datos Masivos: Big Data

Entréguese na Administración da ETSE **como mínimo tres meses antes da data de depósito do TFM**.

1. Datos do/a Alumno/a			
Nome: Raúl Alberto Barrantes Pampillo	**DNI:** 58057453R		
Enderezo: Rua Fernando III O Santo, 18, 7 Esquerda	**Localidade:** Santiago de Compostela		
Provincia: A Coruña	**C.P.:** 15701	**Teléfono:** 644926715	**Correo-e:** raul.pampillo@gmail.com

1. Requisitos

O alumnado deberá estar matriculado na materia do TFM para poder presentar a proposta de anteproxecto.

1. Datos do Traballo de Fin de Máster

Título: Social Big Data y enfermedad mental: aplicación automática de nuevas técnicas discursivas a colecciones de redes sociales en contextos de enfermedades mentales.

Titor/a: David Losada Carril — **Correo-e:** david.losada@usc.es

Cotitor/a: Patricia Martín Rodilla — **Correo-e:** patricia.martin.rodilla@udc.es

Cotitor/a: — **Correo-e:**

(o titor debe ser PDI doutor. En caso de TFMs desenvolvidos en empresa o(s) supervisor(es) do TFM na empresa deben figurar aquí como cotitor(es) do TFM e debe figurar un titor académico (PDI) –artigo 5 do reglamento-)

Áreas de coñecemento (do titor académico): Ciencia de la Computación e Inteligencia Artificial

Departamento (do titor académico): Electrónica y computación

A persoa que asina e cos datos que se indican solicita á Comisión Académica do Máster a aprobación do anteproxecto que se acompaña.

Santiago de Compostela, 23 de marzo de 2020

O/A alumno/a	Vº e Pr. O/A Titor(a)/Cotitor(a) do traballo
	LOSADA CARRIL DAVID ENRIQUE - 33297020N Firmado digitalmente por LOSADA CARRIL DAVID ENRIQUE - 33297020N Fecha: 2020.03.23 09:11:21 +01'00' MARTIN RODILLA PATRICIA - 76129001J Firmado digitalmente por MARTIN RODILLA PATRICIA - 76129001J Fecha: 2020.03.22 15:18:00 +01'00'
Asdo: Raúl Alberto Barrantes Pampillo	Asdo:

Á atención da Comisión Académica do Máster universitario en Tecnologías de Análisis de Datos Masivos: Big Data

1. DESCRICIÓN DO ANTEPROXECTO

(Adaptar o tamaño do documento segundo sexa necesario.)

Título

Social Big Data y enfermedad mental: aplicación automática de nuevas técnicas discursivas a colecciones de redes sociales en contextos de enfermedades mentales.

Introdución

En las redes sociales se encuentran habitualmente textos escritos por personas diagnosticadas con enfermedades mentales. En base a esto se pueden encontrar factores discursivos en común. Se pueden usar métricas y patrones discursivos como factor diferencial e indicativo de textos producidos por personas diagnosticadas con depresión, comparándolos con población no diagnosticada. Estos patrones o pistas lingüísticas pueden indicar una patología mental que podrían facilitar o advertir el diagnóstico de tales enfermedades.
Primeramente, dentro de los textos en general se puede medir la coherencia por medio de la frecuencia de palabras dentro de las oraciones, mediante lo cual se asigna un índice de coherencia discursiva [3]. Por otro lado, existen palabras o tópicos que son más comunes dentro de los encontrados en los foros de redes sociales relacionados a salud mental [2][11][12][13][14]. Utilizando elementos procedentes de este tipo de análisis se plantea evaluar factores en común entre los textos escritos por los diagnosticados por depresión.
Tomando como base del trabajo diferentes de textos de prueba y varios estudios que se han hecho al respecto, se pretende implementar y comparar una serie de métricas de coherencia discursiva con el fin de determinar qué elementos puede ser diferenciadores de las personas que sufren trastornos psicológicos como la depresión [9].

Motivación

Para el desarrollo del proyecto se van a usar como base colecciones de texto para estudios de depresión [1], se trata de publicaciones (posts o comentarios) realizadas por un conjunto de personas en la red social Reddit.
Se propone usar varios métodos para el análisis de los textos. Por un lado, Latent Semantic Analysis, construye un espacio 'semántico' con los términos más asociados entre sí y que conduce a una representación matricial [6]. Y por medio de su implementación en Python+Gensim se representa esa información por medio de un modelo de espacio vectorial.
Por otro lado el marco de referencia proporcionado por CODRA (COmplete probabilistic Discriminative) analiza la estructura sintáctica de los textos y podría plantear las relaciones y la coherencia que existe en un texto en forma de árbol [7].
Una vez planteada la estructura semántica de las oraciones se pretende usar métricas de coherencia, con el fin de etiquetar los datos recopilados para su posterior análisis o entrenamiento. Las métricas de coherencia indican la calidad de un documento, y de este modo se puede intentar estudiar la generación de texto y su entendimiento [5][15]. Estas métricas o patrones pueden indicar lo coherente que es un texto dadas sus palabras y frecuencias de las mismas. Asimismo también se considerarán medidas de correlación entre oraciones del texto de una forma sintáctica y semántica [3].
Se pueden percibir atributos lingüísticos generales en común en textos de pacientes diagnosticados con depresión tales como expresiones relacionadas con conductas afectivas, términos relacionados con soporte social, o aspectos relacionados con elementos de anonimización [2][11][12][13][14]. Además se tendrá en cuenta la legibilidad de los textos, por medio de medidas aplicadas por las fórmulas Flesch Reading Ease y el nivel de grado Flesch-Kincaid [4].

Obxectivos

- Buscar patrones que puedan ser usados para detectar índices de depresión.
- Desarrollar software que represente colecciones de textos disponibles en el área de análisis de lenguaje y trastornos depresivos.
- Desarrollar software que analice en términos de coherencia discursiva los textos escritos por distintos grupos de sujetos y visualice los resultados apropiadamente.

Relación cos coñecementos e competencias proporcionadas polo Máster

- Modelos de predicción
- Técnicas de extracción de características en el texto
- Procesamiento de lenguaje natural
- Minería de textos

Plan de Traballo

1. Definición del problema y anteproyecto.
2. Limpieza de datos
3. Análisis y exploración de los datos.
4. Cómputo de métricas de coherencia discursiva.
5. Evaluación de los modelos.
6. Análisis de las métricas evaluadas.
7. Desarrollo de software que integre todos los resultados del proyecto.

Bibliografía

[1] Losada David E., Crestani Fabio, A Test Collection for Research on Depression and Language Use. In *International Conference of the Cross-Language Evaluation Forum for European Languages,* pages 28-39*,* 2019.

[2] Munmun De Choudhury, Sushovan De, Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity, In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media,* 2014.

[3] Lapata Mirella, Barzilay Regina, Automatic Evaluation of Text Coherence: Models and Representations. In *IJCAI'05: Proceedings of the 19th international joint conference on Artificial intelligence,* pages 1085–1090, 2014.

[4] McNamara Danielle S., Louwerse Max M., Graesser Arthur C. , Coh-metrix: Automated cohesion and coherence scores. To predict text readability and facilitate comprehension. In *Behavior research methods, instruments, & computers: a journal of the Psychonomic Society,* 36(2):193-202, 2005.

[5] Li Jiwei and Jurafsky Dan, Neural Net Models of Open-domain Discourse Coherence. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 198-209, 2017.

[6] Deerwester Scott, Dumais Susan T. , Furnas George W., and Landauer Thomas K. Indexing by Latent Semantic Analysis, In *Journal of the American Society for Information Science*. 41(6):391-407*,* 1990.

[7] Joty Shafiq, Carenini Giuseppe, Ng Raymond T. CODRA: A Novel Discriminative Framework for Rhetorical Analysis. In *Computational Linguistics, Volume 41, Issue 3,* pages 385–435 2015.

[8] Elsner Micha, Austerweil Joseph, and Charniak Eugene, A Unified Local and Global Model for Discourse Coherence. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 436–443, 2007.

[9] Barzilay Regina, Lapata Mirella, Modeling Local Coherence: An Entity-Based Approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 141–148, 2008.

[10] Petersen Casper, Lioma Christina, Simonsen Jakob Grue, Larsen Birger, Entropy and Graph Based Modelling of Document Coherence using Discourse Entities: An Application to IR. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, pages 191–200, 2015.

[11] Al-Mosaiwi, M., & Johnstone, T., Corrigendum: In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation. In *Clinical Psychological Science*, 7(3), 636–637, 2018.

[12] Ríssola Esteban, Losada David, Crestani Fabio. Discovering Latent Depression Patterns in Online Social Media. In *10th Italian Information Retrieval Workshop*, pages 13-16, 2019.

[13] Morales Michelle Renee, Scherer Stefan, Levitan Rivka, A Cross-modal Review of Indicators for Depression Detection Systems. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 1–12, 2017.

[14] Dan Iter, Jong H. Yoon, and Dan Jurafsky, Automatic Detection of Incoherent Speech for Diagnosing Schizophrenia. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 136–146, 2018.

[15] Duran Nicholas, Bellissens Cedrick, Taylor Roger, McNamara Danielle. Quantifying Text Difficulty with Automated Indices of Cohesion and Semantics. In *Proceedings of the 29th Annual Meeting of the Cognitive Science Society,* pages 233-238, 2007.

[16] Lal Alice, Tetreault Joel, Discourse Coherence in the Wild: A Dataset, Evaluation and Methods. In *Proceedings of the SIGDIAL 2018 Conference*, pages 214–223, 2018.

[17] Rehurek Radim and Sojka Petr. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks,* pages 46-50, 2010.

Observacións.

FASES DO TRABALLO E ESTIMACIÓN TEMPORAL

Un traballo de fin de máster suporá 450 horas de traballo total do alumno (incluindo titorías, traballo aútonomo e avaliación).

Dedicación semanal prevista (en horas/semana): 25 horas por semana

FaseEstimación temporal (en semanas)	18
1. Definición del problema y anteproyecto.	2
2. Limpieza de datos	2
3. Análisis y exploración de los datos.	**2**
4. Cómputo de métricas de coherencia discursiva.	**3**
5. Evaluación de los modelos.	**3**
6. Análisis de las métricas evaluadas.	**3**
7. Desarrollo de software que integre todos los resultados del proyecto.	**3**