Trabajo Fin de Máster: Título del Trabajo

Trabajo Fin de Máster

Nombre del Alumn@

Trabajo de investigación para el

Máster en Lenguajes y Sistemas Informáticos

Universidad Nacional de Educación a Distancia

Dirigido por el

Prof. Dr. D. Nombre Direct@r

Marzo 2015

Agradecimientos

Agradecimientos si procede.

Resumen

En el presente trabajo se propone una arquitectura para la detección del machismo en la red social Twitter. En la actualidad, el abuso online se ha convertido en un gran problema, especialmente por la anonimidad an y interactividad de la web que facilita el incremento y permanencia de este tipo de abusos. Se trata de un campo en el que ha aumentado la producción científica enormemente durante este mismo año y donde se han desarrollado competiciones con gran participación por parte de la comunidad científica. A lo largo del trabajo, se presenta el ciclo completo para la recolección de datos, preprocesamiento y construcción del sistema de clasificación. Se desarrolla un sistema... Se evalua... Los resultados demuestran ... Finalmente, se identifican algunos problemas y líneas de trabajo futuras.

Completar cuando se avance en el trabajo

Abstract

Breve resumen del trabajo realizado y de los objetivos conseguidos en inglés.

Índice general

Índice de Figuras

Índice de Tablas

Capítulo 1

Introducción

1.1. Motivación

Motivación del trabajo a realizar.

El primer párrafo de cada sección no se indenta, los siguientes sí. Las referencias a un artículo del estado del arte se ponen así: El trabajo bla bla.

La presente plantilla es meramente orientativa. Tanto los capítulos como las secciones de cada capítulo son simplemente una guía para ayudar al alumno en el proceso de escritura del trabajo final del máster, pero en ningún caso supone que los trabajos tengan que tener forzosamente esta estructura. La estructura del trabajo debe ser acordada por el alumno y su director/res, así como adecuarse a la temática abordada.

1.2. Propuesta y objetivos

Que se ha realizado en el trabajo de fin de master, cuales eran los objetivos y breve resumen de los resultados obtenidos. Texto de prueba 3.

1.3. Estructura del documento

Breve descripción de los capítulos del trabajo.

Capítulo 1. Introducción. Este capítulo introduce los principales motivos que han llevado a la realización de este trabajo, así como la problemática y el estado actual de la disciplina. Por último, se presentan las diferentes contribuciones del trabajo realizado.

Capítulo 2. Estado del arte. Este capítulo describe en mayor detalle la disciplina que nos ocupa, presentando su origen y su historia hasta el presente. Se muestran las técnicas actuales más utilizadas para resolver las tareas más relevantes del tema abordado, así como sus debilidades.

Capítulo 3. Sistema/Método/Caso de Estudio propuesto. En este capítulo se describe en profundidad el sistema/método o caso de estudio propuesto.

Capítulo 4. Evaluación. Este capítulo describe la metodología utilizada para evaluar la propuesta realizada, a la vez que presenta los resultados obtenidos al evaluar el método propuesto en diferentes tareas y sobre colecciones de evaluación de distintos dominios.

Capítulo 5. Discusión. Este capítulo analiza y discute en profundidad los resultados obtenidos en la evaluación presentada en el capítulo anterior.

Capítulo 6. Conclusiones y trabajo futuro. Este capítulo recopila las diferentes conclusiones extraídas del trabajo realizado, y propone algunas líneas de trabajo futuro.

Capítulo 2

Estado del arte

El presente capítulo tiene como objetivo presentar al lector la detección del lenguaje machista en redes sociales. Para ello, se realizará una revisión de los trabajos más relevantes en la tarea de detección de lenguaje abusivo y machista, en los que se analizarán los orígenes de esta tarea, las soluciones técnicas y las aportaciones más relevantes.

2.1. Detección de lenguaje o discurso del odio (*hate speech detection*)

La detección del lenguaje machista o sexista está muy relacionada con la detección del lenguaje o discurso del odio en redes sociales. Existen numerosos trabajos donde se intenta detectar distintos tipos de lenguaje del odio, entre ellos el sexismo (HAJIME WATANABE, 2018; Zeerak Waseem, 2016; Georgios K. Pitsilis y Langseth, 2018; Pinkesh Badjatiya, 2017; Steven Zimmerman, 2018; Park y Fung, 2017; Waseem, 2016). El lenguaje del odio se refiere al uso de lenguaje agresivo, violento u ofensivo hacia un grupo específico de personas que comparten una propiedad en común, sea esta propiedad su género, su raza, sus creencias o su religión (Thomas Davidson, 2017). Atendiendo a esta definición, se puede considerar la detección del machismo como un caso particular del discurso del odio. Por ello, es muy interesante realizar una evaluación de los trabajos realizados en esta línea de investigación.

La detección del lenguaje del odio es una linea de investigación muy actual, el primer estudio evaluado data del año 2012 (Guang Xiang, 2012). En

este articulo se emplea un modelo de detección de temas o categorías (*topic modelling*) que explota la concurrencia de palabras para la creación de atributos o *features* que alimentarán un algoritmo de clasificación de aprendizaje de máquina o *machine learning*. En la mayoría de trabajos previos se empleaban soluciones basadas en patrones para la clasificación de tweets. De este modo, este artículo supone un paso muy importante hacia la automatización y a los sistemas basados en algoritmos de *machine learning*. Además, durante la etapa anterior a este artículo, el uso de expresiones coloquiales y soeces en redes sociales hace difícil establecer las fronteras entre el uso de lenguaje ofensivo que no tiene como objetivo despreciar a ningún grupo de personas y el lenguaje del odio (Thomas Davidson, 2017) utilizando patrones extraídos de la utilización del lenguaje.

Durante los últimos tres años, se han sucedido los artículos en la temática y ha aumentado considerablemente la producción científica en este campo. En (Zeerak Waseem, 2016) se aporta el primer corpus de referencia anotados que se utilizará posteriormente en (Waseem, 2016; Georgios K. Pitsilis y Langseth, 2018; Pinkesh Badjatiya, 2017; Steven Zimmerman, 2018; Park y Fung, 2017). Está compuesto por 16.000 *tweets* etiquetados para mensajes sexistas, racistas o sin contenido ofensivo. En este primer trabajo, se sientan las bases de las soluciones aplicadas en el resto de artículos, se utilizan atributos como los *unigramas, bigramas, trigramas* y *cuatri-gramas* y un algoritmo de regresión logística para la clasificación.

En el artículo desarrollado por el mismo autor (Waseem, 2016) se propone una solución similar pero se amplía el corpus en 4033 *tweets* y se utiliza una plataforma de *crowdsourcing* para anotar los mensajes. Achacan el empeoramiento de los resultados al posible sesgo que se produce en (Zeerak Waseem, 2016) ya que los *tweets* solo fueron etiquetados por los autores únicamente.

En el resto de artículos que evalúan su propuesta utilizando el corpus desarrollado por (Waseem, 2016), se utilizan redes neuronales para la tarea de clasificación y, en algunos, en la etapa de preprocesamiento. En la solución propuesta por (Steven Zimmerman, 2018) se aplican redes neuronales convolucionales (*CNN, Convolutional Neural Network*) para codificar el texto y extraer los atributos que se utilizarán para el clasificador final, basado también en CNNs. Esta técnica permite tener en cuenta la posición de la palabra (su contexto) para extraer los atributos de cada *tweet*. Esta

misma idea junto con el uso de redes neuronales recurrentes (*RNN, Recurrent Neural Network*) se utiliza en (Pinkesh Badjatiya, 2017) para obtener los atributos en la etapa de procesamiento. En ambos artículos se consiguen mejorar los resultados alcanzados por (Waseem, 2016).

En (Georgios K. Pitsilis y Langseth, 2018) se propone un modelo basado en RNNs para abordar el problema. Además se explora la idea de utilizar atributos como la tendencia al racismo o sexismo utilizando el historial de los usuarios. Se demuestra como el uso de este tipo de atributos mejora notablemente los resultados. Esta misma idea se utiliza en (Despoina Chatzakouy, 2017) donde se detectan cuentas agresivas estudiando al usuario y su red de seguidores.

En todos los artículos revisados anteriormente, se trata el problema como una clasificación múltiple donde el texto se puede clasificar según las etiquetas racismo, sexismo o ninguno. Sin embargo, se podría resolver el problema con un doble clasificador, el primero clasifica si el texto contiene lenguaje abusivo o no y el segundo realizaría la tarea de clasificar en contenido sexista o racista (Park y Fung, 2017).

Un desafío importante en la detección del lenguaje del odio en redes sociales es la separación entre lenguaje ofensivo y el lenguaje que incita o promueve el odio. Davidson (Thomas Davidson, 2017) aporta un corpus etiquetado de 25.000 *tweets* para diferenciar entre estos 2 tipos de lenguaje. En su trabajo, se propone un modelo similar a (Waseem, 2016) donde se ponen de manifiesto las dificultades de esta solución para tener en cuenta el contexto de las palabras. De este modo, si se utilizan palabras que pueden expresar odio (por ejemplo, "*gay*") en un contexto positivo, hay muchas probabilidades de que el sistema detecte odio en el texto. Los resultados serán mejorados posteriormente en (HAJIME WATANABE, 2018) donde se ampliará el número de*features* y se utilizará un algoritmo basado en árboles de decisión para la tarea de clasificación.

2.2. Detección de la misoginia

La misoginia se define según la RAE como *"Aversión a las mujeres"* (RAE, b). El machismo, por contra, se refiere a "Actitud de prepotencia de los varones respecto de las mujeres" o "forma de sexismo caracterizada por la prevalencia del varón" (RAE, a). Si bien estos dos términos tienen

matices distintos, tienen como denominador común la discriminación de las mujeres debido a su sexo. De hecho existen trabajos donde se manifiesta que la misoginia se manifiesta lingüísticamente mediante la exclusión, discriminación, hostilidad, trato de violencia objetificación o cosificación sexual (Maria Anzovino y Rosso, 2018; E. Fersini y Anzovino, 2018a). Muchas de estas señales textuales de misoginia serían aplicables del mismo modo al machismo (Aranbarri, 2014; Giraldo, 1972).

Durante este último año, se ha llevado a cabo la competición IberEval 2018 donde una de las tareas era la detección automática de la misoginia (https://amiibereval2018.wordpress.com/, 2018) (AMI, *"Automatic Misogyny Identification"*). En esta tarea se pronone la tarea de identificar la misoginia en *tweets* en español e inglés. En total, participaron 11 equipos distintos de 5 países para la detección en inglés mientras que para la detección en castellano participaron un total de 8 equipos (E. Fersini y Anzovino, 2018b). Los artículos publicados para esta tarea en castellano resultan de gran interés pues guarda una relación importante con el presente trabajo.

Para la tarea de clasificación, la mayoría de los equipos utilizaron Máquinas de Vectores de Soporte (SVM, *Suppor Vector Machines*) y métodos combinados de aprendizaje (EoC, *Ensemble of Classifiers*). Las técnicas basadas en SVMs fueron utilizadas por (Canós, 2018; Endang Wahyu Pamungkas y Patti, 2018; Nina-Alcocer, 2018). Los equipos (Resham Ahluwalia y Cock, 2018; Elena Shushkevich, 2018; Simona Frenda y y Gomez, 2018; Han Liu y Cocea, 2018) aplicaron técnicas EoC mientras que en (Goenaga y Perez, 2018) se exploraron soluciones basadas en redes neuronales.

Las soluciones aportadas por (Canós, 2018; Endang Wahyu Pamungkas y Patti, 2018) obtuvieron la mejor tasa de aciertos para la detección de la misoginia en castellano. El modelo propuesto por (Canós, 2018) utiliza *features* basadas en la vectorización de cada tweet utilizando la medida tf-idf (*term frequency - Inverse document frequency*) y, posteriormente, se utiliza un modelo SVM con nucleo lineal para la tarea de clasificación. Esta solución tan sencilla alcanza los mejores resultados para *tweets* en castellano pero empeora considerablemente para el inglés.

Una idea interesante explorada en (Endang Wahyu Pamungkas y Patti, 2018) es el uso de un léxico auxiliar que contenga palabras que se encuentran con frecuencia en textos sexistas. Este léxico fue desarrollado en un trabajo italiano (De Mauro, 2016). En este modelo se utiliza como clasificador, un

modelo basado en SVM con núcleo lineal para el castellano y núcleo radial para el inglés. En este caso, se alcanza la máxima tasa de aciertos en inglés y en español.

2.2.1. Ejemplo subsección

Capítulo 3

Sistema/Método/Caso de Estudio propuesto

En este capítulo se describe en profundidad el sistema/método o caso de estudio propuesto. La organización de este capítulo dependerá sustancialmente del trabajo abordado.

3.1. Ejemplo sección

3.1.1. Ejemplo subsección

Capítulo 4

Evaluación

Este capítulo describe la metodología utilizada para evaluar el sistema/método o caso de estudio propuesto, a la vez que presenta los resultados obtenidos en la evaluación de las diferentes tareas y sobre colecciones de evaluación de distintos dominios. Algunos ejemplos de secciones pueden ser estos:

4.1. Metodología de evaluación

4.2. Métricas de evaluación

4.3. Colecciones de evaluación

4.4. Resultados

Capítulo 5

Discusión

Este capítulo analiza y discute en profundidad los resultados obtenidos en la evaluación presentada en el capítulo anterior. La estructura de este capítulo dependerá del tema del trabajo y de la estructura del capítulo anterior.

5.1. Ejemplo sección

5.1.1. Ejemplo subsección

Capítulo 6

Conclusiones y trabajo futuro

Este capítulo recopila las diferentes conclusiones extraídas del trabajo realizado, y propone algunas líneas de trabajo futuro. Las siguientes secciones son las que suelen contener este tipo de capítulos, aunque pueden variar dependiendo del tema del trabajo.

6.1. Conclusiones

6.2. Trabajo futuro

Bibliografía

Bibliografía

[Aranbarri2014] Aranbarri, Garazi Urdangarin. 2014. Cosificación de las adolescentes en las redes sociales digitales, pág 43. Master's thesis, Universidad del País Vasco.

[Canós2018] Canós, Jose Sebastián. 2018. Misogyny identification through svm at ibereval 2018. En *IberEval 2018*.

[De Mauro2016] De Mauro, T. 2016. Le parole per ferire. En *Internazionale (2016)*.

[Despoina Chatzakouy2017] Despoina Chatzakouy, Nicolas Kourtellisz, Jeremy Blackburnz Emiliano De Cristofaro Gianluca Stringhini Athena Vakaliy. 2017. Mean birds: Detecting aggression and bullying on twitter. En *WebSci*.

[E. Fersini y Anzovino2018a] E. Fersini, P. Rosso y M. Anzovino. 2018a. Overview of the task on automatic misogyny identication at ibereval 2018. En *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*.

[E. Fersini y Anzovino2018b] E. Fersini, P. Rosso y M. Anzovino. 2018b. Overview of the task on automatic misogyny identification at ibereval 2018. En *IberEval 2018*.

[Elena Shushkevich2018] Elena Shushkevich, John Cardiff. 2018. Classifying misogynistic tweets using a blended model: The ami shared task in ibereval 2018. En *IberEval 2018*.

[Endang Wahyu Pamungkas y Patti2018] Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile y Viviana Patti. 2018. Exploiting lexical knowledge for detecting misogyny in english and spanish tweets. En *IberEval2018*.

[Georgios K. Pitsilis y Langseth2018] Georgios K. Pitsilis, Heri Ramampiaro y Helge Langseth. 2018. Detecting oensive language in tweets using deep learning. En *Department of Computer Science Norwegian University of Science and Technology*.

[Giraldo1972] Giraldo, Octavio. 1972. El machismo como fenómeno psicocultural. En *Revista Latinoamericana de Psicología*.

[Goenaga y Perez2018] Goenaga, A. Atutxa, K. Gojenola A. Casillas A. Daz de Ilarraza N. Ezeiza M. Oronoz A. Perez y O. Perez. 2018. Automatic misogyny identification using neural networks. En *IberEval 2018*.

[Guang Xiang2012] Guang Xiang, Bin Fan, Ling Wang Jason I. Hong Carolyn P. Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. En *CIKM, 2012, Maui, HI, USA*.

[HAJIME WATANABE2018] HAJIME WATANABE, MONDHER BOUAZIZI, TOMOAKI OHTSUKI. 2018. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. En *2018 IEEE*.

[Han Liu y Cocea2018] Han Liu, Fatima Chiroma y Mihaela Cocea. 2018. Identification and classication of misogynoustweets using multi-classier fusion. En *IberEval 2018*.

[https://amiibereval2018.wordpress.com/2018] https://amiibereval2018.wordpress.com/. 2018. Automatic misogyny identification, ibereval 2018.

[Maria Anzovino y Rosso2018] Maria Anzovino, Elisabetta Fersini y Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. En *Springer International Publishing AG, part of Springer Nature 2018*.

[Nina-Alcocer2018] Nina-Alcocer, Victor. 2018. Ami at ibereval2018 automatic misogyny identification in spanish and english tweets. En *IberEval 2018*.

[Park y Fung2017] Park, Ji Ho y Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. En *Proceedings of the First Workshop on Abusive Language Online*.

[Pinkesh Badjatiya2017] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. En *2017 International World Wide Web Conference Committee (IW3C2)*.

[RAEa] RAE, Definición Machismo. http://dle.rae.es/srv/search?m=30&w=machismo.

[RAEb] RAE, Definición Misoginia. http://lema.rae.es/dpd/srv/search?key=misoginia.

[Resham Ahluwalia y Cock2018] Resham Ahluwalia, Evgeniia Shcherbinina, Edward Callow Anderson Nascimento1 y Martine De Cock. 2018. Detecting misogynous tweets. En *IberEval 2018*.

[Simona Frenda y y Gomez2018] Simona Frenda, Bilal Ghanem y Manuel Montes y Gomez. 2018. Exploration of misogyny in spanish and english tweets. En *IberEval 2018*.

[Steven Zimmerman2018] Steven Zimmerman, Chris Fox, Udo Kruschwitz. 2018. Improving hate speech detection with deep learning ensembles. En *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

[Thomas Davidson2017] Thomas Davidson, Dana Warmsley, Michael Macy Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. En *ICWSM 2017*.

[Waseem2016] Waseem, Zeerak. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. En *Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Science*.

[Zeerak Waseem2016] Zeerak Waseem, Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. En *Proceedings of NAACL-HLT 2016*, páginas 33–41.

Apéndice A

Publicaciones

Publicaciones derivadas del trabajo realizado.

Todo list

v, Completar cuando se avance en el trabajo