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

Octubre 2018

Agradecimientos

Agradecimientos si procede.

Resumen

En el presente trabajo se propone una arquitectura para la detección del machismo en la red de microblogging Twitter. En la actualidad, el abuso online se ha convertido en un gran problema, especialmente por el anonimato y la 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

Con el rápido crecimiento de las redes sociales, la comunicación entre personas de diferentes culturas en todo el mundo se ha convertido mucho más directa y sencilla. Esto provoca un gran aumento de los "ciber" conflictos entre las personas que utilizan con frecuencia este tipo de plataformas. Con millones de contribuciones e información generada diariamente por los usuarios de este tipo de herramientas, resulta impracticable y poco escalable realizar una política manual para detectar el abuso y el machismo. Pese a esto, empresas como Facebook han anunciado planes para contratar varios miles de empleados encargados de moderar el contenido de la plataforma (Quartz, 2017). Pese a dedicar muchos esfuerzos y recursos, las grandes compañías como Twitter encuentran gran cantidad de dificultades para afrontar el problema (Atlantic, 2016) debido a la gran cantidad de posts que no pueden ser mediados por sus moderadores. Además, han impulsado fuertes iniciativas para responder a las críticas recibidas por no atajar el problema con la suficiente contundencia. Twitter, por ejemplo, ha aplicado politicas para prohibir el uso de sus plataformas para atacar a personas o grupos sociales (Twitter: (Twitter, 2018)). La importancia de este problema junto con la gran cantidad de información generada por los usuarios hace necesaria la creación de sistemas y herramientas que puedan automáticamente detectar el contenido inapropiado en redes sociales.

Un nuevo estudio realizado en EEUU (Duggan, 2017) sostiene que el 41 % de personas encuestadas había sufrido personalmente algún tipo de discriminación o acoso online, de las cuales el 18 % había sufrido algún tipo

de acoso grave, por ejemplo debido a su género (8 %). De igual modo, las mujeres tienen más de el doble de probabilidades de sufrir acoso debido a su género. (ESTE PARRAFO PROVOCA QUE LA SEPARACION ENTRE PARRAFOS CAMBIE)

La importancia del problema radica en la posibilidad de que un abuso verbal en redes sociales acarree eventualmente un acto de violencia física. De hecho, no es inusual que contenidos machistas hacia las mujeres sean trasladados a acciones violentas. Por ejemplo, algunos estudios sociales como (Fulper y Rowe, 2014) demuestra la existencia de una correlación entre el número de violaciones y el número de tweets machistas por estado en USA. Esto sugiere que las redes sociales pueden ser utilizadas como detector de violencia machista.

Amnistía internacional, publicó recientemente un estudio donde denuncia este hecho (International, 2017). En el reporte, se explica cómo para muchas mujeres Twitter es una plataforma donde la violencia y al abuso contra ellas florece, en la mayoría de los casos sin ninguna consecuencia. Según este informe, Twitter está fallando como empresa a la hora de respetar los derechos de la mujer en línea. En lugar de reforzar las voces de las mujeres, la violencia y el abuso que experimentan en la plataforma hace que las mujeres se autocensuren a la hora de postear, limiten sus interacciones e incluso les hace abandonar Twitter por completo. De este modo, la violencia y el abuso que muchas mujeres experimentan en Twitter tiene un efecto perjudicial en su derecho a expresarse en igualdad, libremente y sin miedo.

Todo lo expuesto justifica sin duda la realización de este trabajo, en el que se propone una arquitectura para la detección automática del machismo. Todos los estudios listados justifican la necesidad de detectar y filtrar de un modo automatizado el contenido que incita o promueve el machismo. En concreto, el lenguaje machista o sexista, ocupa gran parte de este discurso en sitios webs como Twitter. Mientras que en la mayoría de las plataformas el uso de este tipo de lenguaje está prohibido, el tamaño de estas redes hace imposible controlar todo el contenido que generan. Para la realización de este proyecto, ...

Completar cuando se avance en el trabajo

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, datando el primer estudio evaluado en el 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. Utilizando estos métodos, el uso de expresiones coloquiales y soeces en redes sociales hace más complicado 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).De este modo, este artículo supone un paso muy importante hacia la automatización y a los sistemas basados en algoritmos de *machine learning*.

Durante los últimos tres años, se han sucedido diferentes artículos en la temática aumentando considerablemente la producción científica en este campo. En (Zeerak Waseem, 2016) se aporta el primer corpus de referencia anotado 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 en 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. Seún los autores, el empeoramiento de los resultados puede deberse 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 en la etapa 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 al sexismo sirviéndose del 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 detecta 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 el 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 considerar 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, sin embargo, se define como "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 expone 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 propone la labor de identificar la misoginia en *tweets* en español e inglés. En total, participaron once equipos de cinco países distintos para la detección en inglés, mientras que para la detección en castellano participaron un total de ocho 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 acierto 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*). Posteriormente, se emplea un modelo SVM con núcleo lineal para la etapa de clasificación. Esta solución tan sencilla alcanza los mejores resultados para *tweets* en castellano, pero empeora considerablemente para *tweets* en inglés.

Una idea interesante, explorada en (Endang Wahyu Pamungkas y Patti, 2018), es el uso de un léxico auxiliar que contenga palabras que se encuentren con frecuencia en textos sexistas. Este léxico fue desarrollado en un trabajo italiano (De Mauro, 2016). En dicho estudio, 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 acierto en inglés y en español.

2.3. Clasificación de textos

El procesamiento del lenguaje natural (NLP) tiene como objetivo fundamental el desarrollo de métodos que permitan a los computadores realizar tareas relacionadas con el lenguaje humano, como la comunicación o el procesamiento de textos.

La principal diferencia del NLP con el resto de líneas de investigación relacionadas con el análisis de datos o la inteligencia artificial es la necesidad de un conocimiento del lenguaje en todas sus aplicaciones. Elementos clave del lenguaje como la fonética, la fonología, la morfología, la sintaxis, la semántica, la pragmática y la discursiva son esenciales en cualquier técnica de procesamiento del lenguaje.

Una de las áreas más importantes de investigación relacionadas con el NLP es la clasificación de textos o documentos. De un modo general, se conoce como clasificación automática a la tarea de asignar una o varias categorías predefinidas sobre una colección de instancias a clasificar. Del mismo modo, la clasificación de textos se puede entender como aquella tarea en la que un documento o texto es etiquetado como perteneciente a un determinado conjunto. Este tipo de técnicas se utilizan para un gran número de aplicaciones:

- Indexación para sistemas de recuperación de información
- Detección de *spam*
- Identificación del lenguaje
- Análisis de sentimientos
- Organización de documentos
- Desambigüación del sentido de las palabras
- Filtrado de textos

Formalmente, el problema se define como un texto o documento d que puede pertenecer a un conjunto fijo de clases $C = \{c_1, c_2, ..., c_i\}$. La salida del sistema como predicción la clase $c \in C$.

Para resolver el problema de la clasificación de textos existen dos enfoques principales: uno basado en reglas y otro mediante algoritmo de clasificación supervisado.

Los sistemas basados en reglas utilizan patrones predefinidos por un experto para crear un conjunto de pautas mediante la combinación de palabras u otros atributos. En este tipo de arquitecturas, la precisión puede ser alta siempre que estas reglas estén cuidadosamente seleccionadas por un experto. Sin embargo, dichos sistemas resultan muy costosos de construir y mantener.

El aprendizaje supervisado se construye sobre un conocimiento a priori. Se debe disponer de un conjunto de documentos de ejemplo para cada una de las categorías consideradas. Después de una etapa de entrenamiento, el sistema queda ajustado de modo que, ante nuevos ejemplos, el algoritmo es capaz de clasificarlos en alguna de las clases existentes. Para este tipo de sistemas se utilizan distintos modelos de clasificadores: *Naive Bayes*, *Regresión logística*, *SVM*, *redes neuronales*, etc.

Para construir cualquier clasificador de textos o documentos es necesario seguir los siguientes pasos:

- Extraer los atributos o *features* necesarias para realizar una representación fiel del texto y que permita la utilización de un algoritmo de clasificación
- Desarrollar procedimientos por los cuales los documentos puedan ser clasificados automáticamente dentro de categorías.
- Evaluar la calidad de la clasificación en relación a algún criterio.

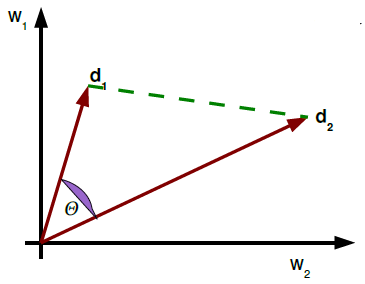
2.3.1. Representación textual

La representación del texto es un paso fundamental para el procesamiento automático de textos. Una representación fiel al contenido del documento, que incluya la información necesaria para extraer conocimiento útil, será clave para el desarrollo de una arquitectura con un rendimiento adecuado. En este proceso, se han de tener en cuenta las especificaciones de los algoritmos que se empleen a continuación.

En esta fase, se definen todos los atributos utilizados en el paso posterior por el algoritmo de clasificación. Los atributos seleccionados o generados a partir de los originales serán los que marquen el éxito de la arquitectura completa. La elección del algoritmo de clasificación para los pasos posteriores in-

fluirá de un modo mucho menos significativo. Por ejemplo, en (Nina-Alcocer, 2018) y (Endang Wahyu Pamungkas y Patti, 2018) se utiliza el mismo algoritmo de clasificación pero los resultados son muy diferentes debido a los atributos utilizados.

Un modelo de representación muy utilizado se conoce como modelo de representación vectorial. Mediante esta representación, los documentos se modelan como vectores dentro de un espacio euclídeo. De este modo, se pueden aplicar operaciones de distancia entre vectores, como indicador de su cercanía según el contenido textual. En la siguiente imagen se muestra un ejemplo en dos dimensiones:



En este caso, se tendría un vocabulario con únicamente dos rasgos w_1 y w_2 que conforman el espacio en el que se encuentran los documentos o textos d_1 y d_2. De este modo, se pueden emplear medidas de distancia, como la distancia euclídea o la distancia coseno, para comparar ambos documentos.

Utilizando este modelo, un texto quedará representado como una combinación lineal de vectores, donde cada coeficiente representa la relevancia de cada rasgo en el contenido del texto, calculado con una función de pesado. Para un texto d, un vocabulario de tamaño n: $\vec{d} = t_1 j\vec{t_1} + ... + t_n j\vec{t_n}$. Para el cálculo de la relevancia de cada rasgo $t_n j$, se utilizará una función de pesado. Una de las más utilizadas se conoce como TF-IDF (frecuencia del termino x frecuencia inversa del documento) y se calcularía del siguiente modo:

$$TF - IDF(\vec{t_i}, \vec{d_j}) = f_{ij} \log(\frac{N}{d_f(\vec{t_i})})$$

donde N es la dimensión del corpus (en este caso número de tweets), f_{ij} la frecuencia del término en el documento y $d_f(\vec{t_i})$ el número de documentos (en este caso el tweet) en los que aparece el término.

2.3.2. Clasificación

Como ya se introdujo en apartados anteriores, la clasificación automática de documentos se puede entender como aquella tarea en la que un documento, o una parte del mismo, es etiquetado como perteneciente a un determinado conjunto, grupo o categoría predeterminada.

Los métodos de clasificación supervisados utilizan un conjunto de documentos de ejemplo para cada una de las categorías que presenta la variable objetivo (a clasificar). Estos algoritmos, realizan una etapa de entrenamiento donde se presentan los patrones de ejemplo de modo que ante futuros patrones, el algoritmo será capaz de clasificar en alguna de las clases contenidas en el conjunto de ejemplo. Dentro de este proceso, existen muchas variables que influirán en los resultados del sistema como el tamaño del conjunto de ejemplo, la elección del algoritmo de clasificación o los parámetros de inicialización del mismo.

Se puede ampliar explicación de cada método

Existen numerosos tipos de algoritmos de clasificación, a continuación se indican los más importantes para clasificación textual:

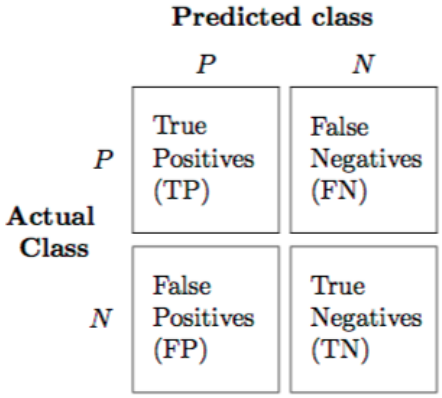
- Naive Bayes: Está basado en la teoría de la decisión de Bayes: la teoría de las probabilidades condicionadas. Por tanto, el problema de la clasificación se reduce al cálculo de las probabilidades a posteriori de una clase dado un documento.
- Arboles de decisión: Se trata de un método que a través de un proceso recursivo de las los atributos de entrada, realiza una representación para clasificar el conjunto de datos presentado.
- Máquinas de vectores de soporte: Estos algoritmos pretenden encontrar una hipersuperficie de separación entre clases dentro del espacio de representación.
- Redes Neuronales: Son un modelo computacional compuesto por elementos ("neuronas") interconectados entre sí que aplican una transformación a los datos para producir una salida. Es posible entrenar una red neuronal para que dada una entrada determinada (un vector de representación) produzca una salida deseada (la categoría a la que corresponde ese documento).
- KNN (K-Nearest Neighbour): Este algoritmo se basa en la aplicación de una métrica que establezca la similitud entre un documento que se quiere clasificar y cada uno de los documentos de entrenamiento.

La clase o categoría que se asigna al documento sería la categoría del documento más cercano según la métrica establecida.

2.3.3. Métodos de evaluación

En la última fase de un sistema de clasificación textual, el modelo se evalúa con un conjunto de datos de prueba en el que se conocen las clases a las que pertenecen sus documentos.

Para la evaluación de los resultados se utiliza comúnmente la matriz de confusión. Se trata de una una herramienta que representa en cada columna el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. En la siguiente imagen se presenta un esquema de la matriz de confusión:



Esta tabla está formada por verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Utilizando estos cuatro componentes se calculas las medidas principales para evaluar los resultados:

- Precisión: representa la fracción de asignaciones correctas frente al total de asignaciones positivas realizadas para esa clase.

$$Precision = \frac{TP}{TP + FP}$$

- Cobertura: representa la fracción de asignaciones positivas respecto al conjunto real de elementos pertenecientes a la clase.

$$Precision = \frac{TP}{TP + FN}$$

- Medida-F: combina las dos medidas anteriores.

$$Medida - F = \frac{2xprecisionxcobertura}{precision + cobertura}$$

2.3.4. Corpus disponibles

A continuación se citan algunos corpus que pueden ser utilizados para la detección de lenguaje del odio en textos:

- IberEval 2018 Automatic Misogyny Identification (E. Fersini y Anzovino, 2018b): Se trata de un corpus etiquetado que contiene campos que denotan si el texto contenido en un tweet tiene un componente sexista. Fue recogido entre el 20-07-2018 y 30-11-2017 donde se recogieron 83 millones de tweets en inglés y 72 millones en castellano. Para el proceso de etiquetado se utilizadon dos pasos: en el primero dos anotadores etiquetaban el conjunto y en el segundo se utilizó una plataforma de crowdsourcing. Finalmente, se etiquetaron 3521 tweets en inglés y 3307 en español para la fase de entrenamiento. En cuanto al conjunto de test, se compartieron 831 tweets en español y 726 en inglés.
- Corpus etiquetado (Zeerak Waseem, 2016): Está compuesto por 16.000 *tweets* etiquetados para mensajes sexistas, racistas o sin contenido ofensivo.

2.4. NLTK: Natural Language Toolkit

NLTK es una librería que define una infraestructura en la que crear programas para el procesado del lenguaje natural (NLP, "Natural language processing") en "Python". Provee la estructura básica para representar datos relevantes para el procesado del lenguaje natural, interfaces para realizar tareas como el etiquetado del discurso (POS, "part-of-speech tagging"), etiquetado sintáctico y clasificación de texto (NLTK, 2018).

Esta librería fue desarrollada originalmente en el año 2001 como parte de un curso de lingüística computacional en la universidad de Pennsylvania. Desde entonces, ha sido desarrollado y mejorado por distintos contribuidores al tratarse de un proyecto libre. Actualmente, NLTK es utilizado en gran cantidad de investigaciones y supone un estándar muy importante para

realizar tareas relacionadas con NLP. Está compuesto por una cantidad importante de módulos que pueden ser invocados desde un programa escrito en Python. En la siguiente figura se recogen los más importantes (Steven Bird y Loper, 2009):

Explicar con más detalle los módulos que utilice en el trabajo

(LA IMAGEN PROVOCA QUE SE DESCUADRE EL DOCUMENTO)

Language processing task	NLTK modules	Functionality
Accessing corpora	nltk.corpus	standardized interfaces to corpora and lexicons
String processing	nltk.tokenize, nltk.stem	tokenizers, sentence tokenizers, stemmers
Collocation discovery	nltk.collocations	t-test, chi-squared, point-wise mutual information
Part-of-speech tagging	nltk.tag	n-gram, backoff, Brill, HMM, TnT
Classification	nltk.classify, nltk.cluster	decision tree, maximum entropy, naïve Bayes, EM, k-means
Chunking regular	nltk.chunk	expression, n-gram, namedentity
Parsing	nltk.parse	chart, feature-based, unification, probabilistic, dependency
Semantic interpretation	nltk.sem, nltk.inference	lambda calculus, first-order logic, model checking
Evaluation metrics	nltk.metrics	precision, recall, agreement coefficients
Probability and estimation	nltk.probability	frequency distributions, smoothed probability distributions
Applications	nltk.app, nltk.chat	graphical concordancer, parsers, WordNet browser, chatbots
Linguistic fieldwork	nltk.toolbox	manipulate data in SIL Toolbox format

2.4.1. 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.

[Atlantic2016] Atlantic, The. 2016. https://www.theatlantic.com/technology/archive/2016/07/twitter-swings-the-mighty-ban-hammer/492209/.

[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*.

[Duggan2017] Duggan, M. 2017. Online harassment 2017. En *Pew Research Center, July 2017*.

[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*.

[Fulper y Rowe2014] Fulper, Rachael, Giovanni Luca Ciampaglia Emilio Ferrara Y. Ahn Alessandro Flammini Filippo Menczer Bryce Lewis y Kehontas Rowe. 2014. Misogynistic language on twitter and sexual violence. En *In Proceedings of the ACM Web Science Workshop on Computational Approaches to Social Modeling (ChASM)*.

[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.

[International2017] International, Amnesty. 2017. Toxic twitter - a toxic place for women. En *Amnesty International Research.*

[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.*

[NLTK2018] NLTK. 2018. https://www.nltk.org/.

[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).*

[Quartz2017] Quartz. 2017. https://qz.com/1101455/facebook-fb-is-hiring-more-people-to-moderate-content-than-twitter-twtr-has-at-its-entire-company/.

[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 Bird y Loper2009] Steven Bird, Ewan Klein y Edward Loper. 2009. *Natural Language Processing with Python.* O'REILLY.

[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.*

[Twitter2018] Twitter. 2018. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

[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
2, Completar cuando se avance en el trabajo
12, Se puede ampliar explicación de cada método
15, Explicar con más detalle los módulos que utilice en el trabajo