Sistema de extracción de características lingüísticas en español para tareas de Procesamiento del Lenguaje Natural

Extracting Spanish Linguistic Features for Natural Language Processing tasks

José Antonio García-Díaz[1]
[1]Universidad de Murcia. Facultad de Informática.
Departamento de Informática y Sistemas
joseantonio.garcia8@um.es

Resumen: El español es el tercer idioma más utilizado en Internet con, aproximadamente, 344 millones de usuarios; este hecho, unido al auge que supuso la Web 2.0. dando a los usuarios un rol principal en la creación de contenido, ha propiciado que el Procesamiento del Lenguaje Natural (PLN) se haya convertido en una de las tecnologías destacadas con aplicaciones en la traducción automática, en sistemas conversacionales o en el desarrollo de filtros de correo no deseado. Sin embargo, en cuanto a recursos disponibles, el PLN en español se encuentra todavía en una fase temprana si lo comparamos con otros idiomas. Además, algunos de esos recursos disponibles han sido desarrollados como traducciones de su equivalente en inglés, por lo que pueden perder características propias del español que no están presentes en el idioma para el que se diseñó el recurso. Por lo tanto, el objetivo de esta tesis doctoral es el desarrollo de un sistema de extracción de características lingüísticas de textos en español, con aplicaciones en diferentes campos del PLN, como la minería de opiniones, detección de plagios o análisis de legibilidad.
Palabras clave: Procesamiento del Lenguaje Natural, Minería de Opiniones, Aprendizaje supervisado, extracción de características lingüísticas

Abstract: Spanish is one of the most popular languages on the Internet with approximately 344 million users; this fact, in conjunction with the rising of the Web 2.0. and the leading role of the users in the creation of content, has leaded Natural Language Processing (NLP) to become one of the outstanding technologies, with applications in machine translation, conversational systems or spam filters. However, some of the available resources are still at an early stage compared to other languages. In addition, some of the tools available are translations of their equivalent in English, so they may lose characteristics of Spanish. Therefore, the objective of this doctoral thesis is the development of a system of extraction of linguistic characteristics of texts in Spanish, which has applications in different fields of the NLP, such as opinion mining, plagiarism detection, or readability analysis.
Keywords: Natural Language Processing, Opinion Mining, Supervised Machine-learning, Linguistic Feature Extraction

1 Introducción

Cada día se generan grandes volúmenes de contenido en Internet. Dentro de la variedad dialéctica que se existe en la red, el español tiene una posición relevante, siendo actualmente el tercer idioma más utilizado en Internet, sólo por detrás del inglés y del chino [1]. Sin embargo, la mayoría de recursos disponibles para poder procesar esa información de manera eficiente están diseñados para el inglés. Por este motivo, la comunidad científica está haciendo verdaderos esfuerzos para crear herramientas de Procesamiento del Lenguaje Natural, o PLN, para el lenguaje español.

Para que un ordenador sea capaz de manejar textos escritos en lenguaje natural hay que codificarlo de manera adecuada. Una es-

[1]https://www.internetworldstats.com/stats7.htm

Lloret, E.; Saquete, E.; Martínez-Barco, P.; Sepúlveda-Torres, R. (eds.) *Proceedings of the Doctoral Symposium of the XXXV International Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)*, p. 32–37

trategia consiste en representar un texto mediante un vector formado por el porcentaje de palabras psico-linguisticamente relevantes, con el objetivo de clasificar palabras que indiquen qué dice el texto, y cómo lo dice. Estos vectores han probado ser efectivos a la hora de clasificar documentos. Por citar algunos ejemplos, se ha aplicado a estudios de autoría (Gaston et al., 2018) o la predicción de resultados electorales (Tumasjan et al., 2010).

Aunque existen herramientas de extracción de características lingüísticas en español, estas no recogen todas las características relevantes del español debido, principalmente, a que son traducciones o adaptaciones de la versión en inglés. Con el ánimo de suplir esta carencia, se pretende el diseño y la implantación un sistema de extracción de características lingüísticas específico para el idioma español con aplicaciones en (1) minería de opiniones, (2) medición del nivel de vocabulario de regiones concretas, (3) tests de legibilidad, (4) estilometría y (5) detección de plagios.

El resto del documento está divido en los siguientes apartados. La sección 2 describe LIWC, el estándar de facto de análisis sintáctico así como se citan trabajos relevantes dentro de la Minería de Opiniones aplicando esta herramienta. La sección 3 detalla la metodología propuesta, haciendo especial hincapié en los prototipos que se están creando. En la sección 4 se listan tres estudios sobre diferentes dominios en los que se está evaluando el prototipo. Por último, la sección 5 sugiere distintos temas de debate que se están planteando durante la realización de esta tesis doctoral.

2 Trabajo relacionado

En la siguiente sección se describe LIWC, un estándar de facto para tareas de análisis lingüístico (Ver sección 2.1) así como investigaciones relacionadas con la Minería de Opiniones (ver sección 2.2) .

2.1 LIWC

LIWC (Tausczik y Pennebaker, 2010) es una herramienta para la extracción de características lingüísticas capaz de analizar un conjunto de textos y generar un vector con los porcentajes de una serie de categorías preestablecidas. Aunque fue originalmente diseñada para el inglés, LIWC cuenta con una versión adaptada al español. Este proceso de traducción fue analizado en (Ramírez-Esparza et al., 2007), donde se identificaron una serie de puntos a mejorar: (1) problemas de traducción entre el inglés y el español, (2) el arbitrario diseño de las dimensiones, (3) diferencias gramaticales no recogidas entre el español y el inglés, (4) conjugaciones verbales insuficientes y, (5) la falta de estudios con fuentes españolas. Además, es importante destacar que LIWC es una herramienta comercial, lo que terminó de motivar el desarrollo de una herramienta libre para la comunidad de PLN en español.

2.2 Minería de Opiniones

El objetivo de la Minería de Opiniones es clasificar si la percepción subjetiva de los usuarios hacía un tema concreto es positiva, negativa o neutra (Esuli y Sebastiani, 2005). En función del nivel de profundidad del análisis deseado, podemos distinguir entre: (1) análisis a nivel de documento, (2) análisis a nivel de sentencia, o (3) análisis a nivel de aspecto. En el análisis a nivel de documento, el texto es clasificado como un todo, devolviendo si la opinión general es positiva, negativa o neutra. En una clasificación a nivel de sentencia el texto se divide en frases y cada una se clasifica de manera individual. Por último, en la clasificación a nivel de aspecto, se trata de clasificar cada aspecto o característica detectada en el documento.

La obtención de la subjetividad se puede realizar mediante (1) Orientación semántica y (2) Aprendizaje computacional. La orientación Semántica consiste en comparar los textos con lexicones compuestos por palabras que reflejan sentimientos, como WordNet-Affect (Strapparava, Valitutti, y others, 2004) o SentiWordNet (Baccianella, Esuli, y Sebastiani, 2010). Por otro lado, los métodos de aprendizaje computacional se basan en entrenar un modelo a partir de un conjunto de instancias ya clasificadas. El modelo resultante debe de ser capaz de replicar el comportamiento humano.

Dentro de las técnicas de aprendizaje computacional, se pueden extraer distintos tipos de características. La técnica más básica de los modelos de aprendizaje supervisado, es decir, del aprendizaje a través de ejemplos, es conocida como Bolsa de Palabras (Bag of Words) y consiste en relacionar la frecuencia de ciertas expresiones con las opiniones del conjunto de entrenamiento. Pese a su simplicidad, el modelo de Bolsa de Palabras fun-

ciona bastante bien; sin embargo, presenta ciertas desventajas. En primer lugar, porque puede sobre-entrenar el modelo, haciéndolo demasiado específico para el conjunto de entrenamiento pero fallando con nuevas instancias. En segundo lugar, porque considera las palabras de manera aislada sin recoger la relación semántica entre el texto, perdiendo información relevante para la clasificación.

Gran cantidad de estudios de minería de opiniones están centrados exclusivamente en documentos en inglés, quizás debido a la falta de recursos en otros idiomas (Martín-Valdivia et al., 2013). Además, un aspecto importante sobre el cual la subjetividad y el análisis de sentimientos requieren mayores esfuerzos está relacionado con el análisis de textos multilíngües. La anotación manual de recursos es una tarea tediosa y costosa, por lo que existen muy pocos corpus y diccionarios para el análisis del sentimiento. Para superar este problema, los investigadores han propuesto métodos para adaptar los recursos existentes y las herramientas para el análisis del sentimiento desarrollado para el idioma inglés para crear recursos en otros idiomas. En este sentido, los lexicones y los corpus anotados se han transformado a nuevos lenguajes utilizando diccionarios bilingües, bootstrapping monolingüe y multilingüe o traducción automática (Balahur y Turchi, 2014). Sin embargo, estos métodos dependen de la disponibilidad y la precisión de los motores de traducción automática.

3 Descripción de la metodología propuesta

En la siguiente sección se describe el sistema de extracción de características lingüísticas en español (Ver sección 3.1) una herramienta de clasificación de corpus en Twitter, para la obtención de corpus de evaluación de la herramienta (Ver sección 3.2) y, por último, la interfaz gráfica de la aplicación (Ver sección 3.3).

3.1 UMUTextStats

UMUTextStats es un sistema de extracción de características lingüísticas diseñado para el español. Al igual que LIWC, este sistema es capaz de extraer un vector formado por los porcentajes de palabras y expresiones que encajan en una serie de características lingüísticas. Sin embargo, se está tratando de resolver las deficiencias que se encontraron en LIWC (Ramírez-Esparza et al., 2007).

UMUTextStats es extensible y permite definir dimensiones a partir de un conjunto de dimensiones abstractas predefinidas, donde destacamos:

- Dimensiones de diccionario. Permite en encontrar expresiones regulares que aparezcan en un determinado catálogo de términos. Esta dimensión también permite indicar contraejemplos. Mediante los contraejemplos es más fácil diseñar una expresión regular sencilla sobre un término, y luego listar las excepciones, como ocurre con el género gramatical.
- Dimensiones basadas en expresiones regulares. Permite, por ejemplo, especificar expresiones regulares para detectar expresiones entrecomilladas, lo que es indicativo del uso de citas textuales o palabras que adquieren algún determinado tono especial.
- Dimensiones basadas en en tipografía. Permite detectar el porcentaje de palabras escritas en mayúsculas, lo cuál puede ser indicio de tono elevado de la voz, característica interesante para la detección de violencia a través de Internet.

Además de estas dimensiones genéricas, se han implementado dimensiones específicas como, por ejemplo, una dimensión para capturar errores gramaticales a partir de la librería PSPell[2] o dimensiones para la detección de verbos, a partir del POSTagger de Stanford[3].

Una ventaja de UMUTextStats frente a otras aplicaciones es que permite operar simultáneamente con distingas versiones del mismo texto. Por lo tanto, algunas dimensiones pueden operar sobre una versión filtrada que facilita la búsqueda de términos en el diccionario, mientras que la versión original se puede utilizar para medir características como el porcentaje de palabras en mayúsculas.

3.2 UMUCorpusClassifier

Con objeto de facilitar el diseño de experimentos para verificar UMUTextSTats, se ha desarrollado también una herramienta de extracción de tweets llamada UMUCorpusClassifier. Esta herramienta permite recolectar

[2]https://www.php.net/manual/en/book.pspell.php
[3]https://nlp.stanford.edu/software/tagger.shtml

corpus de entrenamiento a partir de una cadena de búsqueda y, opcionalmente, una localización geográfica.

Los tweets obtenidos se pueden clasificar de dos maneras. Por un lado, mediante supervisión distante (Go, Bhayani, y Huang, 2009) estableciendo algún tipo de regla automática. Por ejemplo, algunos estudios de la búsqueda de tweets satíricos han partido de la asumpción de que todos los tweets con el hashtag #sarcasm son irónicos (Liebrecht, Kunneman, y van Den Bosch, 2013). Por otro lado, mediante clasificación manual, donde la calidad de la clasificación depende del número de usuarios que clasifican el mismo documento de manera independiente. De esta manera, el sistema potencia cuáles son los documentos que tienen más consenso entre los usuarios descartando los que generen más controversia. Independiente del sistema de clasificación, cada corpus se permite el uso de una escala diferente, aunque por defecto se usa una configuración de cinco niveles: muy positiva, positiva, neutra, negativa, muy negativa y fuera del dominio.

3.3 Interfaz de usuario

La interfaz gráfica de UMUTextStats está integrada con distintas fuentes de donde recoger documentos. En primer lugar, se pueden obtener documentos directamente desde la API de Twitter. En segundo lugar, se pueden subir documentos con los textos en distintos formatos, como CSV, ficheros de texto plano o ficheros comprimidos. En tercer lugar, se pueden comprobar artículos de la Wikipedia a partir de especificar el título. Por último, se ha integrado una comunicación directa con UMUCorpusClassifier. Los vectores de características generados se pueden exportar en diferentes formatos como *JSON*, *CSV*, *HTML* y ficheros *ARFF* para la suite WEKA[4].

Como elemento adicional, la interfaz permite efectuar comparativas con otros modelos. Actualmente, está diseñado para compararse con un modelo de N-Gramas generado a partir de secuencias de palabras o de caracteres.

4 Metodología y experimentos propuestos

En la siguiente sección se describen los experimentos llevados a cabo para la validación de UMUTextStats en diferentes dominios: (1) el estudio de la sátira, (2) infodemiología y (3) el análisis de opiniones sobre economía.

4.1 Sátira

Además de divertida, la sátira es una herramienta constructiva que permite a la sociedad detectar y sobreponerse a sus debilidades. Sin embargo, aunque algunos autores han comparado el periodismo satírico con las noticias falsas, estas difieren en la intencionalidad. Mientras que la sátira pretende crear una versión de la realidad donde nadie espera que sea real, las noticias falsas tienen la intencionalidad de confundir, generar odio, prejuicio o decepción. Debido a la gran capacidad de difusión de las noticias hoy en día, hemos verificado la eficacia de UMUTextStats para entrenar modelos capaces de distinguir entre noticias satíricas y noticias reales ya que, aunque la sátira no es real, no debería de ser considerada contenido pernicioso. Además, la clasificación de la sátira puede ayudar a otras tareas del Procesamiento del Lenguaje Natural, como la Minería de Opiniones, porque el significado implícito de textos satíricos difiere del texto explícito. Consecuentemente, la identificación de contenido satírico nos ayudaría a: (1) diferenciar entre contenido objetivo y divertido, (2) filtrar noticias falsas sin perjudicar el contenido divertido, y (3) identificar contenido que utiliza lenguaje figurado.

Siguiendo esta línea de investigación, se han extraído características lingüísticas a partir de varios corpus encontrados en la bibliografía, además de la recolección de un nuevo corpus balanceado formado por 10.000 tweets, escritos tanto en castellano y en español de México. Para la clasificación de los tweets se ha seguido una estrategia de supervisión distante basando en la presunción de que los tweets son satíricos sólo si provienen de un medio satírico, siguiendo la misma idea que (del Pilar Salas-Zárate et al., 2017) y (Barbieri, Ronzano, y Saggion, 2015). Además, hemos podido comparar nuestro modelo con estos trabajos previos y con un modelo base formado por una Bolsa de Palabras.

Además de las dimensiones genéricas, en este experimento se crearon dos tipos de categorías más. Una categoría específica para extraer características propias de Twitter, como el uso de menciones, hashtags o emoticonos, y otra categoría específica para capturar técni-

[4]https://www.cs.waikato.ac.nz/ml/weka/

cas propias del lenguaje figurativo a partir de la taxonomía especificada en (Roberts y Kreuz, 1994) y que consta de (1) hipérboles, (2) idiotismos, (3) peticiones indirectas, (4) ironía verbal, (5) atenuación, (6) metáforas, (7) preguntas retóricas y (8) símiles.

Los resultados obtenidos han sido positivos mejorando UMUTextStats a los modelos obtenidos en (del Pilar Salas-Zárate et al., 2017) y (Barbieri, Ronzano, y Saggion, 2015). Este trabajo ha sido enviado para su publicación y se encuentra ahora mismo en fase de revisión.

4.2 Infodemiología

La infodemiología es la ciencia que investiga el uso de información disponible en Internet con el fin de mejorar los servicios sanitarios. En este sentido, existen dos grandes tipos de enfoque. Debido al impacto socio-económico causado por las enfermedades infecciosas hemos realizado un conjunto de experimentos con la colaboración de la Universidad de Guayaquil. En concreto, se recolectó y se clasificó un corpus formado por tweets procedentes de Ecuador a partir de palabras clave de virus como el Zika o el Chikungunya. Para estos corpus utilizamos la herramienta UMUCorpusClassifier (Ver sección 3.2) , para recolectar el corpus y para realizar una clasificación manual. Para ello, contamos con 20 estudiantes de la universidad de Guayaquil. que realizaron un total de 51.127 clasificaciones manuales. Como los alumnos clasificaron varias veces los mismos tweets existe cierto consenso en cuanto a las opiniones, y pudimos descartar los tweets que generaban más polémica. Se han llevado ya varios experimentos previos que se han publicado en (García-Díaz et al., 2018a) y (García-Díaz et al., 2018b).

4.3 Economía

En tercer lugar, se ha aplicado este estudio al dominio de la economía. El objetivo de esta tarea era analizar mensajes de usuarios en redes sociales para determinar qué combinación de características lingüísticas podría predecir opiniones positivas y negativas con respecto de la economía. Este estudio se llevado por dos vías. En primer lugar, se realizó un análisis preliminar con un corpus balanceado de 1.000 tweets positivos y 1.000 tweets negativos. Estos tweets fueron clasificados manualmente por personal del laboratorio. En este sentido, el corpus, no tiene tanta aceptación como el otro corpus. Los resultados de este experimento se pueden consultar en (García-Díaz et al., 2018c).

5 Temas específicos a discutir sobre la investigación

UMUTextStats, al igual que LIWC, tiene un sistema arbitrario de dimensiones. Para solucionar el problema, nos hemos puesto en contacto con lingüísticas de la Universidad de Murcia, para asesorarnos en establecer una mejor taxonomía.

Además, se está analizando cuando ciertas expresiones, como "*Nueva York*" deben de ser analizadas como una entidad única o cuando las palabras que lo conforman deben de identificarse en otras categorías. En este sentido, la palabra "*Nueva*" no debería de contabilizar en otras categorías como adjetivos. Sin embargo, si que hay casos donde las palabras si que deben de aparecer en distintas categorías. Para solucionarlo, se ha planteado usar un sistema de reconocimiento de entidades nombradas para tratar estos elementos de manera aislada.

Sobre la herramienta de UMUCorpusClassifier se pretende mejorar el sistema de tweets duplicados. Durante los experimentos se han detectado casos de tweets muy parecidos donde sólo se ha variado alguna coma o símbolo de puntuación. Para esto, se han buscado diferentes herramientas como la propuesta en (Rieck y Wressnegger, 2016), pero no se han implantado todavía.

Agradecimientos

Este trabajo ha sido apoyado por la Agencia Estatal de Investigación (AEI) y el Fondo Europeo de Desarrollo Regional (FEDER) a través del proyecto KBS4FIA (TIN2016-76323-R)

Bibliografía

Baccianella, S., A. Esuli, y F. Sebastiani. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. En *Lrec*, volumen 10, páginas 2200–2204.

Balahur, A. y M. Turchi. 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56–75.

Barbieri, F., F. Ronzano, y H. Saggion. 2015. Is this tweet satirical? a computational approach for satire detection in spanish. *Procesamiento del Lenguaje Natural*, 55:135–142.

del Pilar Salas-Zárate, M., M. A. Paredes-Valverde, M. A. Rodriguez-Garcia, R. Valencia-García, y G. Alor-Hernández. 2017. Automatic detection of satire in twitter: A psycholinguistic-based approach. *Knowledge-Based Systems*, 128:20–33.

Esuli, A. y F. Sebastiani. 2005. Determining the semantic orientation of terms through gloss classification. En *Proceedings of the 14th ACM international conference on Information and knowledge management*, páginas 617–624. ACM.

García-Díaz, J. A., Ó. Apolinario-Arzube, J. Medina-Moreira, H. Luna-Aveiga, K. Lagos-Ortiz, y R. Valencia-García. 2018a. Sentiment analysis on tweets related to infectious diseases in south america. En *Proceedings of the Euro American Conference on Telematics and Information Systems*, página 21. ACM.

García-Díaz, J. A., O. Apolinario-Arzube, J. Medina-Moreira, J. O. Salavarria-Melo, K. Lagos-Ortiz, H. Luna-Aveiga, y R. Valencia-García. 2018b. Opinion mining for measuring the social perception of infectious diseases. an infodemiology approach. En *International Conference on Technologies and Innovation*, páginas 229–239. Springer.

García-Díaz, J. A., M. P. Salas-Zárate, M. L. Hernández-Alcaraz, R. Valencia-García, y J. M. Gómez-Berbís. 2018c. Machine learning based sentiment analysis on spanish financial tweets. En *World Conference on Information Systems and Technologies*, páginas 305–311. Springer.

Gaston, J., M. Narayanan, G. Dozier, D. L. Cothran, C. Arms-Chavez, M. Rossi, M. C. King, y J. Xu. 2018. Authorship attribution vs. adversarial authorship from a liwc and sentiment analysis perspective. En *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, páginas 920–927. IEEE.

Go, A., R. Bhayani, y L. Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12):2009.

Liebrecht, C., F. Kunneman, y A. van Den Bosch. 2013. The perfect solution for detecting sarcasm in tweets# not. En *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity*. New Brunswick, NJ: ACL.

Martín-Valdivia, M.-T., E. Martínez-Cámara, J.-M. Perea-Ortega, y L. A. UreñA-López. 2013. Sentiment polarity detection in spanish reviews combining supervised and unsupervised approaches. *Expert Systems with Applications*, 40(10):3934–3942.

Ramírez-Esparza, N., J. W. Pennebaker, F. A. García, R. Suriá Martínez, y others. 2007. La psicología del uso de las palabras: Un programa de computadora que analiza textos en español. *Revista mexicana de psicología*, 24(1):85–99.

Rieck, K. y C. Wressnegger. 2016. Harry: A tool for measuring string similarity. *The Journal of Machine Learning Research*, 17(1):258–262.

Roberts, R. M. y R. J. Kreuz. 1994. Why do people use figurative language? *Psychological science*, 5(3):159–163.

Strapparava, C., A. Valitutti, y others. 2004. Wordnet affect: an affective extension of wordnet. En *Lrec*, volumen 4, páginas 1083–1086. Citeseer.

Tausczik, Y. R. y J. W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Tumasjan, A., T. O. Sprenger, P. G. Sandner, y I. M. Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. En *Fourth international AAAI conference on weblogs and social media*.