UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

Linguistic features integration for text classification tasks in Spanish

Integración de características lingüísticas para tareas de clasificación de textos en Español

D. José Antonio García-Díaz

2022

UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

Linguistic features integration for text classification tasks in Spanish

Integración de características lingüísticas para tareas de clasificación de textos en Español

Author:
José Antonio GARCÍA-DÍAZ

2022

UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

Linguistic features integration for text classification tasks in Spanish

Integración de características lingüísticas para tareas de clasificación de textos en Español

Author:
José Antonio GARCÍA-DÍAZ

Supervisor:
Dr. Rafael VALENCIA-GARCÍA
Dr. Pedro José VIVANCOS-VICENTE

April, 2022

Acknowledgements

Quiero dedicar esta tesis doctoral a mi director Rafael Valencia García quien, durante todos estos años, me ha tratado más como a un hijo que como un estudiante. Además de tener en cuenta mis opiniones y valorar mis aportaciones, ha sabido dirigirme en todo momento y buscar siempre lo mejor para mí. De él me queda la gran lección de: *Lo bueno (y lo malo) de este trabajo es que todo lo que hagas te beneficia a tí.* Además, esta tesis me ha permitido conocer a grandes profesionales y personas como Salud, Paco, Ángela, Eugenio o Ricardo, de quienes tengo tanto que aprender. También quiero dedicar este trabajo a la empresa VÓCALI SISTEMAS INTELIGENTES S.L. y a mi codirector Pedro José Vivancos Vicente.

En lo personal, no tengo sino palabras de afecto a todas aquellas personas que me han acompañado y ayudado durante todo este camino. No sólo a mis padres, Alfonso y Florentina, a mis hermanos, Alfonso y Damián, a mi tía Gloria y a mi primo Pedro (por quien comenzó mi amor por la informática) o a mis sobrinos, Alejandro y Arturo, sino a todas aquellas personas que, aunque no me unen vínculos de sangre, considero que son parte de mi familia. A Trinidad Fresno, por enseñarme el verdadero valor de la amistad, y por ser una persona a quien si todos imitáramos, el mundo sería un lugar mejor. A Rebeca Simón, quien me da su cariño (y paciencia) todos los días, y de quien aprendo a ver el mundo con otros ojos, descubriendo unos colores que hace poco no sabía ni que existían. Tengo también la suerte de contar con amigos maravillosos, como Álvaro (mi compitrueno -me tienes seco, Pepe-), Paco y Paloma (el descubrimiento de 2022), Abel, Andrés y José (¡Menudo cuadro!), a mis cracks: Tomás, Prip, Esmeralda, Rubén, Gonzalo, Patri, Juanan y Bea (Este es el año de retomar las buenas costumbres), Iria (la voz de mi conciencia) y Abel (quien tiene la misma maldición que yo con el humor), Miriam (¡qué ganas de acompañarte a estrenar la maleta!), a Yiya y Libertad (¡Somos la vieja guardia!), a Iris y a Mariano (deseando compartir más cosas con vosotros), a Adrián (aguardando el castigo un millón), a toda mi familia musical con José Carlos, Antozo, David, Jaime, Daniel, Mariano, Fernando, Carlos, Jandro y, en especial, a Mario Galindo (Mario, nunca te olvidaremos); a mi familia de fulares, con Carlos Manuel (mi alma es tuya), Juan Luís, Tere, Jesús, Luís, Mayka y María. También quiero agradecer muchísimo a mis compañeros de laboratorio, con Fran, Ginés, Jésica, Juan, José, Laura y Camilo. ¡Cómo echo de menos aquellos cafés que el confinamiento nos quitó!. También a mis compis de Vialan e Imaginanet, de quien tanto he aprendido. Ana y Manolo (de vuestro Za), Juanjo, Pablo, Ramón, Jesús, Amaluye, Carlos García y Barreiro, Ana y Paco, y a Lorena y Jesús; y a mis compañeros de carrera Juanjo, Rubén, Queen, Tocayo, Agustín, Laura, Belén (mi BFF), Guadalupe, Miguel Ángel, Juanico, Antoñico, y Ginés.

Por último, quiero dedicar mi cariño a Profi y Bubu, mis queridos compañeros de innumerables paseos y miembros de todo derecho de mi familia.

UNIVERSIDAD DE MURCIA

Abstract

Linguistic features integration for text classification tasks in Spanish

by José Antonio GARCÍA-DÍAZ

The state-of-the-art concerning automatic document classification relies on language models. These models learn the complexity of human language from massive datasets using unsupervised learning. In these models, words are represented as vectors. These vectors are similar to words that share meaning and context. Once learnt, these vectors can be rearranged for solving classification tasks such as Hate-speech detection, Sentiment Analysis or Author Analysis, among other tasks. However, despite they achieve spectacular performance, the resulting models are difficult to interpret. Furthermore, in languages such as Spanish, the latest state-of-the-art language models are not immediately available as they need to be created specifically for them. We argue that the usage of a limited set of hand-made linguistic features produce models that are easier to interpret, with competitive performance, and that generalise better. Moreover, linguistic features and embeddings can be combined by applying different strategies such as knowledge integration or ensemble learning, improving the performance of both. The de facto tool for extracting linguistic features in Spanish is LIWC (Linguistic Inquiry Word Count). This tool encodes texts as linguistic features organised into a set of relevant psycho-linguistic categories. However, the translation of LIWC to Spanish had some drawbacks, being the most relevant one the loss of certain aspects of Spanish during the translation process. Accordingly, we present two Natural Language Processing tools designed for the Spanish language. The main one is UMUTextStats, a linguistic extraction tool that captures many aspects of linguistics, such as phonetics, lexis, morphosyntax, stylometry, semantics, or pragmatics, among others. The other tool is UMUCorpusClassifier that eases the compilation and annotation of linguistic corpora. In addition, we present a exhaustive validation of UMUTextStats with the publication of four articles in high impact journals and the participation in several international workshops. It is worth mentioning that this evaluation is not limited to Spanish but we have evaluated a subset of these features in English and low-resource languages.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	**A**rtificial **I**ntelligence
BoW	**B**ag **o**f **W**ords
BERT	**B**idirectional **E**ncoder **R**epresentations from **T**ransformers
CNN	**C**onvolutional **N**eural **N**etwork
LIWC	**L**inguistic **I**nquiry and **W**ord **C**ount
LSTM	**L**ong **S**hort **T**erm **M**emory
MI	**M**utual **I**nformation
MLM	**M**asked **L**anguage **M**odel
MLP	**M**ulti **L**ayer **P**erceptron
NLP	**N**atural **L**anguage **P**rocessing
NSP	**N**ext **S**entence **P**rediction
RNN	**R**ecurrent **N**eural **N**etworks
RMSE	**R**oot **M**ean **S**quiare **D**eviation
RoBERTa	**R**obustly **O**ptimized **BERT** **P**retraining **A**pproach
PCA	**P**rincipal **C**omponent **A**nalysis
PoS	**P**art **o**f **S**peech
SVM	**S**upport **V**ector **M**achine
TF–IDF	**T**erm **F**requency – **I**nverse **D**ocument **F**requency
XML	e**X**tensive **M**arked **L**anguage

Chapter 1

Resumen

1.1 Motivación

El Procesamiento del Lenguaje Natural es la rama de la Inteligencia Artificial y de la Lingüística que facilita la comunicación entre las personas y los computadores a través del lenguaje natural. El Procesamiento del Lenguaje Natural está ahora mismo en un punto disruptivo. Esto se debe principalmente a los avances en el campo del Big Data, que facilita poder trabajar con volúmenes de datos previamente inimaginables y a los avances en el campo del deep-learning, con modelos basados en mecanismos de atención que pueden codificar información en base al contexto. Estos dos factores han facilitado la generación de nuevos modelos del lenguaje pre-entrenados sobre un gran volumen de datos que se pueden adaptar con poco esfuerzo a solucionar distintas tareas, tales como la traducción automática o el resumen automático de texto.

Esta tesis doctoral se centra en la tarea de la clasificación automática del texto, que consiste en etiquetar documentos a partir de una serie de categorías predefinidas. Esta tarea sirve, por ejemplo, para anotar sentimientos sobre un conjunto de textos y así poder determinar la polaridad subjetiva de un texto, tal y como lo haría una persona [88]. Para poder llevar a cabo este tipo de tareas, es necesario desarrollar métodos prácticos para codificar el lenguaje humano. El estado de la técnica de la representación computarizada de textos se basa en modelos basados en Transformers que generan representaciones vectoriales de palabras y sentencias. Estos vectores se aprenden a partir de aprendizaje no supervisado, empleando estrategias tales como *Masked Language Model* (MLM) o *Next Sentence Prediction* (NSP). Esta técnica ha conseguido alcanzar resultados espectaculares en distintas tareas relacionadas con el lenguaje humano. Sin embargo, los modelos resultantes de estas técnicas son complejos y difíciles de interpretar [85].

Otra manera de representar el lenguaje natural es mediante características lingüísticas que capturen ciertos rasgos significativos del lenguaje. Con estos rasgos es posible capturar *qué* dice un texto y *cómo* lo dice [91].

La principal hipótesis de esta tesis doctoral es que las características lingüísticas pueden combinarse con modelos basados en Transformers. Esta combinación tendría dos ventajas. Por un lado, se mejoran los resultados alcanzados por cada conjunto de manera individual y, por otro lado, se dota de cierta interpretabilidad a los modelos resultantes.

En concreto, en esta tesis doctoral se describe el desarrollo y evaluación de la herramienta UMUTextStats[1]. Esta herramienta permite obtener estadísticas de un texto manuscrito organizadas en una serie de características relevantes. Aunque algunas de estas características pueden aplicarse a distintos idiomas, UMUTextStats ha sido diseñada para el español. UMUTextStats está inspirada en LIWC [91], que es la herramienta de facto para extraer características lingüísticas. En sus orígenes, LIWC se diseñó para el inglés, aunque esta herramienta ha sido adaptada a distintos idiomas tales como chino [53], portugués [8], francés [74], alemán [66], o neerlandés [12], por citar algunos ejemplos. También hay una versión en español [79]. Sin embargo, se identificaron ciertas limitaciones durante su evaluación ya que ciertos rasgos específicos del español son ignorados como, por ejemplo, el género gramatical.

Entre otras aplicaciones, las características extraídas con UMUTextStats se pueden aplicar a tareas de clasificación automática de documentos. De hecho, la herramienta ha sido validada en distintos dominios, tales como el análisis de sentimientos basado en aspectos [33], la identificación de misoginia [42] y el discurso de odio [43], el perfilado de autores [34], o la identificación de la sátira [37]. Todos estos trabajos han sido publicados en revistas de alto impacto, estando cuatro de estos trabajos presentados como un compendio en esta tesis doctoral. Además, las características lingüísticas han sido evaluadas en workshops internacionales, tales como IberLEF, SemEval, o FIRE. En estas competiciones, las características lingüísticas han sido evaluadas tanto de manera aislada como combinadas con modelos del estado de la técnica, consiguiendo resultados competitivos en casi todas las tareas.

Como contribución adicional de esta tesis doctoral, se ha desarrollado la herramienta UMUCorpusClassifier[2] [47], que sirve para compilar y etiquetar corpus lingüísticos de manera automática o semiautomática.

1.2 Objetivos y metodología

En esta tesis doctoral se analizan dos hipótesis principales. Por un lado, que la inclusión de características lingüísticas capaces de capturar rasgos de los autores mejora el desempeño de los sistemas de clasificación automática (RH1) y que, por otro lado, estas características lingüísticas mejoran la interpretabilidad de los modelos resultantes (RH2).

[1] https://umuteam.inf.um.es/umutextstats/
[2] https://umuteam.inf.um.es/corpusclassifier

Para llevar a cabo estas hipótesis hemos definido las siguientes objetivos:

- **OB1**. Obtención de una taxonomía de las diferentes características lingüísticas del español.
- **OB2**. El desarrollo de la herramienta UMUTextStats y del léxico relacionado con cada característica dentro de la taxonomía.
- **OB3**. El desarrollo de la herramienta UMUCorpusClassifier para la compilación y anotación de corpus en español.
- **OB4**. Validación de la herramienta UMUTextStats en diferentes dominios.
- **OB5**. Recopilación y anotación de corpus lingüísticos en español para realizar tareas de clasificación automática de textos en diferentes dominios.

1.3 Resultados

Cumplir con los objetivos marcados en esta tesis doctoral ha permitido publicar nuestras propuestas y resultados en revistas científicas de alto impacto, además de poder participar en congresos y conferencias internacionales. Los principales resultados obtenidos se presentan en esta tesis doctoral como compendio. Se incluye, además, una lista con todas las publicaciones derivadas de esta tesis doctoral en la Tabla 4.1.

A continuación, se detallan los principales resultados de cada uno de los artículos del compendio.

1.3.1 Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in Latin America

La primera publicación [33] está relacionada con la tarea de análisis de sentimientos basado en aspectos sobre el dominio de la Infodemiología. La Infodemiología es el proceso de extraer información relacionada con la salud pública en Internet con el objetivo de mejorar los sistemas públicos de salud [25]. En concreto, el objetivo de este trabajo fue el de catalogar publicaciones en redes sociales de la ciudadanía sobre asuntos relacionados con enfermedades infecciosas.

La primera tarea a este respecto fue la de compilar un corpus en español relacionado con la salud. El corpus fue compilado de Twitter utilizando la herramienta UMUCorpusClassifier. En concreto, se analizaron tres enfermedades infecciosas: el Zika, el Dengue y el Chikunguña. La anotación del corpus fue llevada a cabo por estudiantes de la Universidad de Guayaquil (Ecuador), que realizaron un total de 51 127 anotaciones, recibiendo cada documento una media de más de 6 anotaciones. Una vez anotado el corpus, evaluamos su calidad con

distintas métricas. Por ejemplo, obtuvimos un 0.6864 de coeficiente de acuerdo entre anotadores (*Krippendorff's Alpha*). Los documentos con menos consenso entre los anotadores fueron descartados, generando una versión definitiva del corpus con 10 843 documentos etiquetados como positivos, 10 843 como negativos, and 7 659 como neutrales.

La segunda tarea fue la de obtener los aspectos relacionados con las enfermedades infecciosas. Para ello, se desarrolló una ontología del dominio. Esta ontología incluye y relaciona conceptos de enfermedades infecciosas tales como síntomas, riesgos de la salud, métodos de transmisión o medicamentos.

La tercera tarea fue la de entrenar y evaluar distintos modelos de aprendizaje computacional para generar un sistema automático de extracción de análisis de sentimientos. En este sentido, asumimos que cada documento tenía un único sentimiento debido a la breve longitud de los mismos. Los modelos de análisis de sentimientos se crearon a partir de las características lingüísticas de UMUTextStats y de word embeddings pre-entrenados en español. Nuestro mejor resultado fue utilizando únicamente las características lingüísticas, obteniendo un *accuracy* de 55.3%. Estos resultados mejoraron, de manera aislada, el resto de las características evaluadas.

La cuarta tarea fue la de asociar los sentimientos a los aspectos. Para ello, obtuvimos los conceptos de la ontología que aparecían en cada documento de manera explícita. A continuación, le sumamos a cada concepto el sentimiento basado en el porcentaje de salida del modelo hacia cada una de las clases y del valor de TF (Term–Frequency). Luego, medimos la distancia entre cada uno de estos conceptos con el resto de los conceptos de la ontología, y ponderamos el valor en función de la distancia entre ambos conceptos. Por ejemplo, si en un documento marcado como positivo aparece explícitamente el término *aspirina*, a este concepto se le suma un valor positivo, así como se suma ese valor ponderado según distancia a los conceptos relacionados con aspirina, tales como medicinas, o síntomas que se tratan con aspirinas. El resultado de este proceso iterativo es que teníamos por cada concepto de la ontología el grado en el que los ciudadanos lo consideraban positivo, negativo o neutro.

La quinta tarea fue la de diseñar una interfaz web donde se visualizan estos conceptos y el sentimiento anotado. Esta interfaz dispone de una serie de gráficas que permite especificar intervalos definidos de tiempo, filtrar por conceptos de la ontología, o bien por zona geográfica.

1.3.2 Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings

En la segunda publicación [42], se presenta un estudio centrado en la identificación de misoginia en redes sociales.

La primera tarea consistió en compilar un corpus en español sobre la misoginia. Este corpus balanceado se llama Spanish MisoCorpus-2020 y tiene 3 841 documentos catalogados como misóginos. El Spanish MisoCorpus-2020 se distribuye completo o bien dividido en tres bloques. El primer bloque está centrado en identificar violencia hacia mujeres con puestos socialmente relevantes (VARW). El segundo bloque se centra en identificar las diferencias del comportamiento misógino en textos compilados en España con textos compilados en América Latina (SELA). El tercer bloque contiene documentos con rasgos genéricos relacionados con la misoginia, tales como el uso de estereotipos o el descrédito (DDSS). Este corpus fue compilado usando la herramienta UMUCorpusClassifier y cada documento fue anotado varias veces por cada uno autores del artículo.

La segunda tarea consistió en entrenar y validar modelos de aprendizaje computacional para la identificación de la misoginia como un problema de clasificación binaria. Nuestra propuesta se basó en combinar las características lingüísticas con sentence embeddings de fastText en español [50]. Nuestro mejor modelo obtuvo un *accuracy* de 85.175%.

Además de para entrenar los modelos de aprendizaje, las características lingüísticas se emplearon para la interpretabilidad de los resultados. Para ello, calculamos la ganancia de información (*Information Gain*) normalizada. Como esperábamos, las características lingüísticas relacionadas con el lenguaje ofensivo resultaron ser muy relevantes para la identificación de la misoginia. También encontramos una correlación positiva entre la misoginia con el género gramatical. Esto es así porque en el español los adjetivos masculinos y femeninos suelen tener significados diferentes y, en muchas ocasiones, los adjetivos femeninos tienen connotaciones peyorativas hacia las mujeres, mientras que sus equivalentes masculinos resaltan virtudes de los hombres. Otras características relevantes identificadas están relacionadas con la estilometría y con características de la categoría de corrección y estilo, lo que sugiere que el uso correcto del lenguaje es relevante a la hora de construir sistemas de detección de contenido misógino.

Además, evaluamos nuestros métodos con corpus propuestos en tareas internacionales relacionadas con la identificación de misoginia, tales como AMI 2018 [28] o HatEval 2019 [11], obteniendo muy buenos resultados y mejorando los resultados obtenidos por los participantes en las tareas de clasificación binaria.

1.3.3 Psychographic traits identification based on political ideology: An author analysis study on Spanish politicians' tweets posted in 2020

La tercera publicación [34] consistió en la compilación y evaluación de un corpus relacionado con dos tareas de análisis de autores: el perfilado de autores y la atribución de autores. El dominio sobre el que evaluamos este trabajo tiene que ver con la política. Este dominio se seleccionó porque, en general, la gente es reacia a

seguir el consejo y directrices de partidos políticos que son de otra ideología. Esto puede llevar a que los ciudadanos tomen ciertas decisiones de manera irracional, poniendo en riesgo su vida y las de las personas de su entorno durante situaciones de emergencia o crisis. Por tanto, además de evaluar sólo rasgos demográficos tales como la edad o el sexo, añadimos un rasgo psicográfico basado en la ideología política. Este rasgo está medido en dos ejes: binario y multiclase.

Para llevar a cabo este trabajo, compilamos el PoliCorpus-2020, un corpus formado por tweets escritos por políticos de España durante 2020. En una primera fase, compilamos cerca de 250 000 tweets de un total de 385 políticos, de los que disponíamos información acerca de su sexo biológico y su año de nacimiento. El espectro político lo anotamos en base al partido político al que están afiliados y a la percepción que tiene la ciudadanía española de la ideología de cada partido. En una segunda fase, eliminamos todos aquellos tweets que no estaban escritos en español o que claramente eran fragmentos de titulares de periódicos. En una tercera fase, seleccionamos los tweets más representativos de cada usuario. Para ello, agrupamos los tweets según el mes en el que fueron escritos y dentro de cada mes los ordenamos de manera alterna en base a una serie de tópicos seleccionados a mano. Una vez ordenados, seleccionamos de manera secuencial un tweet de cada mes y de cada tópico, hasta llegar a un mínimo de 120 tweets por usuario. En una cuarta fase, el corpus fue anonimizado para dificultar la identificación de los usuarios.

Una vez el corpus fue compilado, evaluamos diferentes modelos de clasificación basados en redes neuronales para solucionar las dos tareas propuestas. Estos modelos probaban las características lingüísticas de UMUTextStats y diferentes modelos basados en word y sentence embeddings. Los resultados obtenidos para la tarea de perfilado de autores fueron muy prometedores. Sin embargo, el corpus compilado tiene el sesgo de que todos los autores son políticos. Para comprobar si los resultados obtenidos eran generalizables, compilamos otro corpus con textos de usuarios que no eran políticos y comprobamos cuanto se degradaban los resultados cuando evaluábamos con los modelos previamente entrenados. En este sentido observamos que los modelos formados por varios conjuntos de características eran más robustos.

Con respecto a la tarea de atribución de autores, los mejores resultados los obtuvimos combinado las características lingüísticas con un modelo basado en Transformers, obteniendo una F1-score de 29.336%.

Por último, analizamos la correlación de las características lingüísticas con cada tarea y etiqueta. En general, los resultados que obtuvimos indican que las características lingüísticas son buenos indicadores para la identificación de la ideología política tanto en binario como en multiclase. De hecho, identificamos que las características basadas en morfosintaxis eran más efectivas en tareas de

perfilado de autores, mientras que las características de estilometría eran más efectivas para la atribución de autores. Además, estas características mejoraron los modelos que estaban basados únicamente en Transformers.

1.3.4 Compilation and evaluation of the Spanish SatiCorpus 2021 for satire identification using linguistic features and transformers

La última publicación presentada como compendio de esta tesis doctoral [37] está relacionado con la identificación de la sátira. Aunque el espíritu de la sátira ha sido desde siempre el hacer una crítica constructiva de la sociedad a través del humor y la burla, es comúnmente confundida con bulos o propaganda, cuyo objetivo es confundir a la gente e influir en la opinión pública. Además, la identificación de la sátira está fuertemente relacionada con la identificación del sarcasmo y de la ironía, así como del lenguaje figurativo en el cual las palabras pierden su sentido literal, dificultando todas aquellas tareas relacionadas con el lenguaje humano.

En concreto, este experimento consistió en compilar y evaluar un corpus lingüístico compuesto por titulares de noticias de prensa satírica y prensa tradicional, tanto de prensa en España como en países de América Latina. Este corpus se llama Spanish SatiCorpus 2021, y es un corpus balanceado formado por 18 207 textos satíricos y otros 18 207 textos no satíricos. Este corpus contiene titulares y otros textos comprendidos entre 2018 y 2021. Además, este corpus trata de solventar algunas de las deficiencias identificadas en otros corpus relacionados. Esta deficiencia tiene que ver en que las noticias satíricas y las no satíricas no siempre tratan sobre los mismos eventos, por lo que los clasificadores automáticos pueden estar sesgados hacia el tópico y no hacia si el texto es satírico o no. Para evitar este problema, se realizó un proceso de emparejamiento, donde se vinculaban aquellos titulares que se referían al mismo hecho. Para ello, se construyó una matriz donde las columnas y las filas eran los identificadores documentos satíricos y no satíricos respectivamente. Cada valor en la matriz era la distancia coseno entre cada par de noticias en base a su similitud textual. Una vez calculada esta matriz, seleccionamos de manera iterativa el par de documentos satírico y no satírico que tuvieran una mayor similitud semántica.

Una vez compilado el corpus, entrenamos varios modelos de clasificación empleando las características lingüísticas de UMUTextStats y distintos tipos de embeddings. Todos estos conjuntos de características fueron evaluados por separado y combinados. Para evaluar como estos conjuntos se complementan mejor, probamos distintas estrategias, tales como métodos de ensemble learning o bien la de integrar cada grupo de características dentro de la misma red neuronal. El mejor resultado se obtuvo con una combinación de BERT y las características lingüísticas dentro de la misma red neuronal, obteniendo un *accuracy* de 97.405%.

Además, comparamos nuestra metodología con datasets en español existentes relacionados con la identificación de la sátira, la ironía o el sarcasmo [9, 72], mejorando la mayoría de resultados.

1.4 Conclusiones y trabajo futuro

Durante esta tesis doctoral hemos mostrado el desarrollo y evaluación de un conjunto de características lingüísticas en español que han probado su efectividad en tareas de clasificación automática. Estas características se pueden extraer con la herramienta UMUTextStats. La idea principal de esta tesis es que estas características se pueden incorporar a modelos de aprendizaje computacional mejorando, por un lado, su desempeño y, por otro lado, su interpretabilidad.

La primera hipótesis se ha demostrado evaluando las características lingüísticas de UMUTextStats en distintos experimentos que se adjuntan como compendio de esta tesis, así como la participación en distintas competiciones internacionales, donde hemos obtenido resultados muy competitivos. Para la segunda hipótesis, obtuvimos para cada experimento la correlación entre las características lingüísticas con las etiquetas de los datasets, analizando el por qué y cuáles son las características más relevantes en dominios como la infodemiología, la identificación de misoginia, el discurso de odio, o perfilado de autores.

Aunque los resultados han sido satisfactorios y prometedores, continuaremos mejorando los diccionarios y el desempeño de cada una de las características lingüísticas, así como traduciendo y adaptando la herramienta a otros idiomas tales como al inglés.

Además, consideramos que las características lingüísticas pueden ser útiles de otras formas. Por ejemplo, para seleccionar mejores particiones de datos de entrenamiento, validación y prueba. En lugar de realizar una muestra aleatoria, las características lingüísticas pueden producir mejores estrategias de muestreo, basadas en la cantidad de palabras, la longitud, los pronombres o las palabras que pertenecen a cierta característica lingüística.

Una limitación de las características lingüísticas es que los vectores generados no recogen el contexto, tal y como pasa en otros modelos estadísticos como la bolsa de palabras o el modelo TF–IDF. Para resolver este problema, proponemos la generación de un limitado conjunto de características lingüísticas extraídas token por token en lugar de todo el documento. Esto nos permitirá utilizar las características lingüísticas como una secuencia y combinarlos con redes neuronales recurrentes y con mecanismos de atención.

Otra prometedora línea de investigación es mejorar la integración de las características lingüísticas con el resto. En los trabajos publicados, hemos evaluado estrategias de integración en la misma red neural o bien aplicando técnicas de

ensemble learning. Estas técnicas de ensemble learning consistían en combinar los resultados obtenidos por cada modelo de manera individual para obtener una nueva predicción aplicando distintas estrategias como calcular la media de las predicciones o bien aplicando algún sistema de voto. En este sentido, evaluaremos *mixture of experts* [23], una estrategia basada en el principio de divide y vencerás, que cubre diferentes regiones de entrada en el espacio del problema con diferentes modelos.

Por último, las técnicas empleadas para evaluar el impacto de las características lingüísticas en cuanto a la interpretabilidad han sido agnósticas al modelo. Sin embargo, consideramos evaluar el desempeño de estas características lingüísticas dentro de una red neuronal. De esta manera, podemos medir su beneficio en modelos que combinan distintos tipos de características.

Chapter 2

Synthesis

2.1 Introduction

Natural Language Processing (NLP) is the part of Artificial Intelligence (AI) and Linguistics that aims at easing the communication between computers and humans using human language. NLP involves different levels, such as parsing, word disambiguation, sentence tagging, machine translation, text analysis, or information retrieval [62]. There are multiple benefits of using NLP. For instance, NLP dismisses social barriers in communication, as tools such as language translators which can be trained with low-resource languages [58]. More friendly user interfaces is another application of NLP, allowing users to use their own voice to communicate with electronic devices. These new user interfaces, which are popular in smart speakers, are inclusive as they reduce technological barriers with elderly people.

From a few years to now, NLP is in a golden moment. Two facts have contributed to this. On the one hand, the rise of big data, focused on dealing with datasets that are incredibly long and that were unmanageable with traditional data-processing methods. On the other hand, the rise of Transformers, a deep-learning architecture which is capable of learning the underlying patterns of language. Nowadays, it is possible to train and evaluate models from large sources of documents, such as books, medical reports, or town hall registration documents, just to name a few.

In this doctoral thesis we focus on automatic classification tasks. This task consists in assigning a series of predefined labels to a set of documents. An example of document classification is Sentiment Analysis (SA), that attempts to determine the subjective polarity of a document [88]. In order to conduct automatic classification tasks, computers need practical ways to represent natural language. The first approaches for categorising human language dealt with statistical models, such as the Bag of Words (BoW) model, that is based on measuring the frequency of the words that make up a text. However, the features extracted with BoW have some important drawbacks, highlighting the curse of dimensionality, that is, if the number of documents is large enough, the vocabulary size increases until the point

that it is impractical in some scenarios. Another important drawback is the lost of word order, which leads to ignore important linguistic phenomena, such as polysemy. These drawbacks have been solved with the usage of state-of-the-art word and sentence embeddings, in which words and sentences are encoded as dense vectors. These vectors are learnt using unsupervised NLP tasks, such as Masked Language Model (MLM) or Next Sentence Prediction (NSP), based on the distributional hypothesis that captures co-occurrence properties of the language [92]. The vectors generated have reached state-of-the-art results in several NLP tasks. However, models learned from word embeddings are complex and hard to interpret as they result in black-box models [85].

Linguistic features is a kind of characteristics that represents documents by means of a vector formed by the percentage of linguistically-relevant traits. These traits capture words and expressions that indicate *what* the text says, and *how* it says it. We argue that linguistic features can be combined with count-based features as well as word and sentence embeddings in order to build better models while providing interpretability of their behaviour.

In this doctoral thesis we describe the development and evaluation of the UMUTextStats tool[1]. A tool for extracting linguistic features. This tool is designed specifically for Spanish, since Spanish is the third most used language on the Internet. It is worth mentioning that there are other linguistic extraction tools available in Spanish, being LIWC [91] the most relevant one. However, as far as our knowledge goes, LIWC does not handle specific linguistic phenomena of Spanish. For instance, Spanish relies on inflection mechanisms to reflect the tense, mood and the person to whom the verb refers. In addition, LIWC is a commercial tool and we aim to provide an open-source tool for the Spanish NLP community.

The linguistic features extracted from UMUTextStats can be applied to automatic text classification tasks. In fact, the UMUTextStats tool has already been applied in several domains and tasks, including aspect-based sentiment analysis in the medical domain [33], the identification of misogyny [42] and hate-speech [43], author profiling tasks for determining demographic and psychographic traits of a set of anonymous users [34], and satire identification [37]. All these works have been published in high-impact journals and four of them are presented as a compendium in this doctoral thesis. This apart, we describe the participation in different international workshops, such as IberLEF, SemEval, or FIRE, in which we evaluate the linguistic features separately and combined with Transformers and traditional machine-learning methods. In these shared-tasks, we have achieved competitive results in almost all of them. These tasks involve hate-speech detection, emotion analysis, humour detection or source-code profiling, among others.

[1] https://umuteam.inf.um.es/umutextstats/

Moreover, the evaluation of UMUTextStats in different domains requires to have different linguistic corpora. Thus, as an extra contribution, we have developed the UMUCorpusClassifier tool[2] [47] that can be used for compiling and annotating linguistic corpora. This tool allows to perform automatic labelling based on some heuristics, or allows administrators to coordinate and supervise teams of human annotators.

2.1.1 Related work

When extracting linguistic features, LIWC [91] is the *de facto* tool. This tool generates a vector with the percentages of a series of pre-established categories from a set of documents. It is worth noting that recently, LIWC has released LIWC-22. However, in order to limit the objectives set at the beginning of this research, our work has focused on the 2010 version.

According to the LIWC's webpage[3], most of its linguistic features are percentages of total words within a text. Other features, however, are raw counts. These features are usually related to summary measures, such as the raw number of words within a document. LIWC contains four summary measures (some of them being present in the previous versions of LIWC), namely analytical thinking, clout, authenticity, and emotional tone. These features are standardised scores converted to percentiles. Next, some details of these features are given. `Analytical thinking` is derived from other linguistic categories and function words. This feature captures how people use *formal, logical, and hierarchical thinking patterns*. When this score is low, this should indicate that they use a more intuitive and personal language. In contrast, high scoring in this feature is linked to reasoning skills. The second summary measure, `clout`, is linked to *social status, confidence, or leadership*. `Authenticity` captures whether people speak spontaneously or not. When this measure is low, the texts usually reflect people that are cautious. On the contrary, high scores in authenticity are usually captured in conversations with close friends or persons with no social inhibitions. Finally, `Tone` combines two features: positive and negative tone. The larger this feature gets, the more positive the tone is.

LIWC has been applied in several studies. It is very popular for conducting author analysis task. For example, LIWC has been applied to authorship attribution [49], narcissism [52], depression and well-being [81, 90], playfulness [77], or decision support [64]. From automatic document classification perspective, LIWC has been explored to discern among satirical and non-satirical headlines from newspapers from Mexico and Spain [72], Sentiment Analysis [61, 7] or deceit detection [3], among others.

[2] https://umuteam.inf.um.es/corpusclassifier
[3] https://www.liwc.app/

2.1.2 Motivation

LIWC was designed for English and it has been translated into other languages such as Chinese [53], Portuguese [8], French [74], German [66], or Dutch [12], just to mention a few.

LIWC has a Spanish version [79]. During its development, two main drawbacks were identified. The major drawback is related to translation problems between English and Spanish. For instance, some grammar differences between both languages are not identified. Besides, the Spanish version of LIWC lacks some of the Spanish verb tenses. The second major drawback is the arbitrary design of the dimensions. Apart from these drawbacks, it is worth mentioning that LIWC is a commercial tool, thus we have an extra motivation for the development of a free tool for the PLN community in Spanish.

2.1.3 Research hypothesis

There are two main research hypotheses discussed in this doctoral thesis. One the one hand, we discuss whether the inclusion of linguistic features that capture linguistic traits of the authors can improve the performance of automatic text classification systems. We put the focus on this study in Spanish, including a wide variety of domains concerning infodemiology, hate-speech, humor, or irony among others. On the other hand, we hypothesise that the inclusion of linguistic features can provide interpretability to the models with a fewer number of features that generalise better than systems built upon Language Models and Transformers.

Accordingly, we define the following research hypotheses concerning the inclusion of linguistic features in automatic classification systems.

- **RH1**. The inclusion of linguistic features improves the performance of automatic text classification systems in Spanish.
- **RH2**. The inclusion of linguistic features can provide interpretability to the models.

2.1.4 Objectives

To accomplish the aforementioned research hypotheses, we define the following objectives:

- **OB1**. Obtaining a taxonomy of the different linguistic features of Spanish.
- **OB2**. The development of the UMUTextStats tool and the related lexicons for each feature within the taxonomy.
- **OB3**. The development of the UMUCorpusClassifier tool for the compilation and annotation of Spanish corpora.

- **OB4**. Validation of the UMUTextStats tool in different scenarios.
- **OB5**. Compilation and annotation of linguistic corpora in Spanish to conduct automatic document classification in different domains.

2.1.5 Thesis structure

The structure of this doctoral thesis is based on a compendium of publications of four research articles that are presented along with the description of the participation in several workshops regarding NLP in which our participation was grounded on the usage of linguistic features from UMUTextStats.

This document is structured into three chapters.

Chapter 2 details all the work produced during this doctoral thesis. Apart from the abstract, the introduction, its motivation and a state-of-the-art subsection with the methodologies and evaluation used, this chapter describes the system architecture of the two tools developed and summarises the experimental results obtained during the validation of the tool, which have given rise to the publications that are presented by compendium and the participation in several international workshops.

Chapter 3 presents the research articles that are attached as the compendium of the doctoral thesis. These research articles are about: (1) an ontology-driven aspect-based sentiment analysis system focused on infodemiolgy [33]; (2) the compilation process and evaluation of the Spanish MisoCorpus 2020, focused on misogyny detection in Spanish [42]; (3) the compilation process of the Spanish PoliCorpus 2020 and its evaluation with two author analysis tasks: an author profiling task to extract demographic and psychographic traits, and an authorship attribution task in order to obtain which the author of a set of anonymous documents [34]; (4) and the compilation process of the Spanish SatiCorpus 2021, which includes satirical headlines and tweets from a wide variety of from Spain and Latin America newspapers [37].

Finally, Chapter 4 contains the conclusions, a summary of all the publications derived from this work and a list of promising future research lines related to the linguistic features and automatic document classification in Spanish.

2.2 State-of-the-art

In this section, different feature engineering techniques related to the linguistic features and novel embeddings are explored. These feature sets refer to the state-of-the-art for conducting automatic document classification tasks.

2.2.1 Feature engineering

Feature engineering is the process of extracting relevant features from raw data. These features can be used as input for building predictive models. Usually, in order to extract relevant features, domain knowledge is applied. A good selection of the features improves the performance of machine learning models. However, there is a large list of feature sets that are too general and can be applied with guarantees in multiple domains.

The feature engineering stage involves (1) the identification of relevant variables and its relationship with the output of the model; (2) the transformation of these variables in order to increase the performance of the model (this step usually involves operations such as normalisation of the data, change its scale or by removing outliers); (3) the extraction of the features, which consists in obtaining the features from the raw data, applying IR techniques, such as TF–IDF; and (4) selection, which consists in discarding irrelevant and redundant features by applying several feature selection algorithms.

2.2.2 Feature sets

The traditional feature sets employed for automatic document classification are described in this Section. These features include statistical features, such as the BoW model and its variants (word-n-grams, char-n-grams, and tf–idf), and word embeddings features. As the UMUTextStats tool is focused on text, we deliberately omit other kind of features, such as contextual or multi-modal features.

Statistical features

The Bag of Words (BoW) model encodes a text as the frequency of each word in the corpus. The BoW model is easy to implement and usually provides good results. For that reason, the BoW model is popular as a baseline model in many NLP tasks. Its major drawback is that it is context-less (that is, it does not consider the surrounding words of a specific word). Therefore, it does not consider linguistic phenomena such as figurative language, polysemy, the presence of typos or other grammatical errors. Moreover, the BoW model is not truly language-independent, as the BoW model works better with non-agglutinative and western languages (for instance, Spanish, English, or Italian). In non-agglutinative language, words usually have a single inflectional morpheme to denote multiple grammatical, syntactic, or semantic features. In agglutinative languages, however, new words are composed by stringing morphemes without changing their spelling. Besides, the BoW model suffers from the *curse of dimensionality*, as the vector size depends on the vocabulary size. For large corpora and documents, the number of features generated makes this model impractical, even it is possible to apply feature reduction techniques such as Principal Component Analysis (PCA). Nevertheless,

despite the aforementioned drawbacks, the BoW model is still a popular approach that achieves good results with small and medium datasets.

There are some strategies for improving the BoW model. For example, instead of handling words in isolation, it is possible to cluster words by distance. This approach is known as the word n-gram model (in fact, the BoW model is the word n-gram model, in which `n` is equal to 1). The benefits of the n-gram model is that it reduces the context-less drawback [93]. For instance, bigrams (`n = 2`) can handle composed words, such as *New York*. However, using n-grams could cause even larger vectors, which hinders the reliability of using linguistic models. In a variant of the n-gram model, characters are used instead of words as linguistic units [57]. The key-advantages of using characters instead of words is that they capture lexical and morphological information, such as punctuation symbols, prefixes or suffixes. Another benefit is that character n-grams are more robust against misspellings, since a word and its misspelling version should share common characters. Moreover, character n-grams behave better in agglutinative languages as they are capable of extracting individual phonemes from compound words.

Both word and character n-grams measure the raw frequency. However, the raw count could be misleading in some scenarios and could bias the models in case of non-informative words such as stop-words. An improved version of the n-grams is the Term Frequency–Inverse Document frequency (TF–IDF) (see Equation (2.2)). In fact, the raw count of terms is just the TF. The TF–IDF algorithm considers how relevant a term inside the whole corpus is, dismissing the importance of terms that appear too often in the texts.

$$TF\text{–}IDF = TF * IDF \tag{2.1}$$

$$TF = number\_of\_occurrences / number\_of\_grams \tag{2.2}$$

$$IDF = \log_2 corpus\_size / documents\_with\_terms \tag{2.3}$$

Embedding based features

The BoW model and its variants are simple and provide competitive results, but they are limited in some scenarios. That is because the BoW model lacks of context and word order. Therefore, the BoW model is not suitable in some NLP tasks, specially in those related to language generation. Moreover, if words are encoded as arbitrary numbers, machine learning algorithms can do wrong assumptions. For example, if the word *dog* is encoded as `10` and *chair* as `20`, a neural model can consider that the concept *chair* is twice the concept *dog*. To avoid this, a popular way to feed individual

words into neural network models is encoding them using one-hot representations. With this representation, each word is represented as a vector of length `N`, being `N` the size of the full vocabulary. For each word, all values are `0` except one. Therefore, all words are orthogonal and there are no dependencies between them. However, with this approach the size of each word is related to the total of words and there is no relationship between the words as we can assume that some words are closer to others as regards of meaning.

Word embeddings solve the aforementioned drawbacks. Word embeddings are dense vectors (instead of the sparse vectors of one-hot encoding representations). The main objective of word embeddings is that words that have similar meaning have similar representation. Word embeddings are considered one of the key breakthroughs of NLP in the last years. The representation of these embeddings are learned using unsupervised tasks.

One of the first proposals for generating word embeddings was word2vec [68]. Specifically, the word2vec model proposed two learning strategies: (1) Continuous bag-of-words (CBOW), in which the order of surrounding words does not influence prediction (similar to the BoW model does). Therefore, the model uses the current word to predict the context. On the other hand, the Skip-gram model weighs heavily nearby context words than the rest of the words.

There are also some alternatives to word2vec. For example, GloVe [70], which makes use of a co-occurrence matrix and neural networks to learn word vectors. FastText [67] is another alternative to word2vec. Contrary to word2vec, fastText learns the word embeddings for each word and surrounding words based on a fixed window size. Then, the values of the embeddings are averaged. Besides, fastText also captures sub-word information. Moreover, fastText can create sentence embeddings by averaging all the words of a text.

One key advantage of word embeddings is that they can be learned from unsupervised tasks and then adjust them to solve specific tasks. This is known as transfer learning. This allows to download pre-trained models of word embeddings based on different algorithms (word2vec, gloVe, fastText) from large corpora. Examples of large corpora in Spanish are the SUC (Spanish Unannotated Corpora) [15], the Spanish Billion Word Corpus [13], or the Spanish Wikipedia. Pre-trained models have two key benefits. They allow to generalise better, as the new models can be aware of words that do not appear during training, and that neural network can converge faster to a solution, as word embeddings already convey general meaning.

In order to conduct automatic document classification, a popular approach to use word embeddings is to average creating a unique vector per document or sentence. These vectors are popularly known as sentence embeddings. Some tools are capable of obtaining sentence embeddings directly, such as doc2vec [59] or fastText

[67]. In case of BERT and similar architectures, the same neural network contains special tokens called classification tokens `[CLS]`. BERT creates a classification token per sentence. However, it is possible to compile documents embeddings from BERT using other alternatives. For example, in [80] the authors evaluate different pooling strategies to average the word embeddings.

However, the representation of a word using these techniques is unique. Therefore, plain word embeddings do not take into account linguistic phenomena such as polysemy. This drawback has been addressed by language models based on Transformers, such as BERT [21] or RoBERTa [60]. These models encode word embeddings taking into account the context, so that the embeddings of the word *date* can be different if it is a verb or a noun. Contextual word embeddings have made a quality leap in many NLP tasks. There are some models adapted to languages such as Spanish [14] or multilingual models [19]. In fact, the *Plan de Impulso de las Tecnologías del Lenguaje*[4] is promoting the development of reusable language models in Spanish. For instance, a model based on RoBERTa has been recently released and trained with a set of documents and web pages crawled by the National Library of Spain [51]. Besides, the trend now is to store in public repositories such as HuggingFace models both general language models and fine-tuned versions in order to solve other tasks such as Question Answering or Named Entity Recognition.

2.3 UMUTextStats

A language can be characterised by a set of features that indicates how words are arranged within a sentence. Linguistic features can refer to grammar aspects of a text, analysing how words and sentences are related. They can capture prosodic features related to stress and intonation, or searching for specific lexicons that can indicate different demographic or psychographic features of the authors. In general, linguistic features creates a model of a language or a specific writing.

UMUTextStats is a linguistic feature extraction system designed for Spanish. Like LIWC, this system is capable of extracting a vector made up of the percentages of words and expressions that fit into a series of linguistic features. However, an attempt is being made to resolve the deficiencies found in LIWC [79]. Some of these drawbacks are shared between the Spanish and the English version of LIWC. First, the arbitrarily design of the linguistic categories and features in which the list of words that belong to a linguistic category was made by a limited number of human annotators. Second, the fact that LIWC is based principally on simple term-count, so that the context of words is not considered. Besides, LIWC does not have categories for all verb tenses. For example, the Spanish LIWC does not contain post-preterites nor past subjunctive.

[4] https://plantl.mineco.gob.es

To solve the aforementioned drawbacks, the UMUTextStats has been designed with a tree-based structure for defining and arranging the linguistic features and categories. We have included several classes for adding not only Dictionary-based dimensions but also patterns with regular expressions and a wide variety of performance errors or specific argot used in social networks. Besides, we have developed a new system for extracting verbs that includes all Spanish's verbal tenses and compound verbs and periphrasis.

2.3.1 Configuration

The design of the UMUTextStats considers software quality attributes concerning maintainability and extensibility. Therefore, the core of UMUTextStats is a configuration file in which the linguistic features are organised within linguistic categories in a tree-based structure.

The UMUTextStats configuration file is an XML file. Listing 2.1 depicts the configuration section of the linguistic feature for capturing expressive lengthening. This linguistic feature is captured with a regular expression `PatternDimension` that captures if the same character occurs more than three times consecutively in a document. As we can observe, we indicate that this feature works with the uncleaned version of the text (`useoriginalinput`).

LISTING 2.1: An example of a feature in the configuration

```
<feature>
    <key>phonetics-expressive-lengthening</key>
    <class>PatternDimension</class>
    <description>Drawing out or emphasizing a verbalized word, giving
        it character</description>
    <pattern>(.)\1{3,}</pattern>
    <useoriginalinput>true</useoriginalinput>
</feature>
```

The dictionaries are extensible. For this, UMUTextStats can create new linguistic features using software classes. These software classes are described below.

- **Dictionary-based features**. This class allows to define new linguistic features that are based in keywords lists. A keywords list is based on regular expressions by default. It is possible, however, to configure this class to disable regular expressions to speed the matching process, achieving O(1) performance.

  Besides, dictionary-based features have others options. For example, it is possible to define counterexamples. The benefit of using counterexamples is that it is easier to define a few general regular expressions, and next, to list the exceptions.

- **Verb-based features**. This class is similar to the dictionary-based features, as it allows verbs to be stored in plain text files. The main difference is that this class uses a custom word separator in order to consider auxiliary verbs as a part of a matching. Besides, the large number of verbs makes the usage of dictionary-based features impractical. Verbs based dimensions are optimised to identify verbs in O(1). For this, this class discards the usage of regular expressions.
- **Sentence-per-Dictionary-based features**. This class obtains how many sentences match certain regular expressions. For instance, with this dimension is easy to get the number of sentences that uses verbs in passive voice.
- **Enclitics-Personal-Pronouns-based features**. This class captures personal pronouns with enclitics. The Spanish pronouns *la*, *lo*, *le*, *los*, *las*, *les*, are enclitics and are used for indicating direct or indirect third-person pronominal object.
- **Perspicuity-based features**. This class obtains the Degrees of Perspicuity according to Flesch-Szigriszt [10].
- **Readability-based features**. This class obtains the readability based on Fernández-Huerta [63].
- **Grammatical-Gender based features**. This class extends the Dictionary-based features. It relies on a list of the basic rules for obtaining the Spanish grammatical gender combined with a list of counterexamples. In addition, this class considers only certain words based on their PoS category. This way, it is easier to discard rare and made up words.
- **STTR-based features**. This class can obtain the Standardised Type/Token Ratio (STTR) [17]. This is the ratio between the total unique words between the total of the words of a text. Besides, for long documents, this class can be configured to obtain the ratio in chunks of `N` words. When using chunks, the output could be the raw count or the standard deviation.
- **Error-Misspellings-based features**. This class gets the number of misspellings using the PSPELL library[5].
- **Error-Misspellings-accents-based features**. This class can detect misspellings based on the wrong usage of accents. For this, this class relies on PSPELL to capture misspellings. Once a misspelling is detected, it checks if the first suggestion of PSPELL contains the same letters.
- **PoS-based features**. This class counts the number of words or expressions that matches certain PoS categories. The PoS categories are calculated using the Stanza Library [78]. The Spanish model of Stanza is built upon the

[5]Based on GNU ASPELL. `http://aspell.net/`

AnCora corpus, which is mainly based on journalist texts. Stanza also considers annotations based on Universal Dependencies, which is a framework that normalises PoS annotations along with grammar, morphological features, and syntactic dependencies.

- **NER-based features**. This class gets the number of words or expressions that matches certain NER categories. Similar to the PoS Tagging Dimension, UMUTextStats relies on Stanza. However, its Spanish model only considers four categories: Person, Location, Organisation, and Miscellaneous. In order to include other categories, it is possible to rely on Dictionary based dimensions, or training a custom NER mdel.
- **Two-or-More-Equal-Words-based features**. This class detects the occurrence of two or more equal words placed next to the other in a text. It is worth noting that this is not strictly an error, but it can indicate the lack of attention on the part of the authors when they are reviewing their texts.
- **Capitalisation-Error-based features**. This class counts how many sentences start with lowercase letter.
- **Pattern-based features**. This class counts how many matches have a custom regular expression. For instance, we can use this class to generate a feature that detects quoted expressions.
- **Typography-based features**. This class allows to detect the number of words written in lower or uppercase. For example, detecting the number of words that are completely written using capital letters could indicate a high tone of the voice.
- **Composite-based features**. This class allows to obtain certain linguistic features using the Composite Pattern. The intent of this pattern is to define a new dimension based on averaging, summing, or calculating the maximum or the minimum.
- **Word-Length-based features**. This class counts the number of words that matches or exceeds a certain threshold. For this, it is possible to configure the word length and the comparative.
- **Word-Average-Length-based features**. This class calculates the average length of all the words within the input.
- **Word-Per-Sentence-based Dimension**. This class calculates the number of words per sentence.
- **Word-Unique-based features**. This class calculates the number of unique words.

- **Syllables-Per-Word-based features**. This class calculates the number of syllables per word.
- **Character-Count-based features**. This class counts the number of specific characters in a text. Characters can be specific using lists. So, it is possible to capture in the same dimension different versions of the same characters, such as quotes, currencies, or brackets.
- **Sentences-Starting-With-the-Same-Word-based features**. This class counts the number of sentences that starts with the same word. This is a custom class to capture certain stylistic errors.
- **Sentences-Starting-With-Numbers-based features**. This class gets the number of sentences that starts with a number. This is considered a bad writing style.
- **Twitter-ReplyTo-based features**. This class determines if certain text (usually, a Tweet from Twitter) is a response to a specific user based on a list of names.

Although each class has specific options that can be specified in the configuration file, there are some common options. Dictionary and pattern-based dimensions, for instance, allow to define a custom separator. The default separator is based on words. However, it is possible to separate the documents by sentences or custom regular expressions. This is useful, for example, to count how many exclamatory sentences are in a text. The configuration also allows to specify if we want raw count or a percentage.

UMUTextStats handles several versions of the same text simultaneously. This is useful because different dimensions can operate on different versions of the text, according to their needs. For example, cleaned versions of the text are more effective when looking up terms in the dictionaries. The uncleaned version of the document, on the other hand, is useful when looking for stylistic errors or counting the percentage of words in capital letters, for instance.

It is possible to provide UMUTextStats with custom cleaned versions of the texts. This is useful, for instance, when the documents have been already pre-processed with a custom tool. If the cleaned version is not provided, UMUTextStats performs a cleaning stage of the texts, similar to the one described for UMUCorpusClassifier (see Section 2.4), that consists in striping blank lines, HTML code, hyperlinks, mentions, and emojis. Besides, each document is transformed to lowercase.

The taxonomy of UMUTextStats is organised in a set of main linguistic categories. These categories are described below.

- **Phonetics (PHO)**. It is the part of linguistics that analyses how humans produce and perceive sounds. The current version of UMUTextStats, which is focused on writing, includes only one feature concerning expressive

lengthening, a linguistic device that consists in repeating some of the letters of a word for emphasis [27]. Table 2.1 contains an example of the expressive lengthening feature.

TABLE 2.1: Some features and examples of Phonetics category

Features	Examples
expressive-lengthening	g*ooooooo*l

- **Morphosyntax (MOR)**. It is the part of linguistics focused on morphology and syntax that studies how words are composed and how sentences are related, respectively.

  Spanish is a highly inflected language. Inflections can denote multiple syntax and semantic meanings that can be used to track stylometric features in author analysis tasks. UMUTextStats divides morphosyntax features into: (1) PoS-based features, that includes adverbs, adjectives, determiners or pronouns, to name but a few; and (2) subword level, that includes features that capture subcomponents of words, such as stems and affixes. This includes features that capture the grammatical gender and number of words.

  This linguistic category has a total of 172 linguistic features. Table 2.2 contains some examples of linguistic features that belong to the morphosyntax category.

TABLE 2.2: Some features and examples of Morphosyntax category

Features	Examples
gender-feminine	gacela, cama, abuela
number-plural	limones, plátanos, luces
affixes-suffixes	acatarrado, tipejo
nouns-common	abogado, martillo
topics-capitals	Madrid, París, Brasilia
topics-countries	España, Francia, Brasil
topics-colours	Azul, verde, amarillo
adjectives-superlative	Rarísimo, clarísimo
adjectives-despective	Flacucho, cabezón, inútil
adverbs-time	Jamás, siempre, nunca
adverbs-mode	Apasionadamente, gratuitamente
adverbs-place	A través, abajo, delante
pronouns-personal	Yo, Tú, Él
pronouns-impersonal	dondequiera, nadie, ningún
prepositions-individual	arriba, dentro, desde
conjunctions-subordinating	que, aunque, pero
verbs-periphrasis	ir(a) + `infinitive`

- **Correction and style (CAS)**. UMUTextStats covers linguistic and stylistic errors. On the one hand, linguistic errors deviate from the accepted rules of Spanish. Examples of linguistic errors are misspellings or the wrong use of

accentuation. On the other hand, stylistic errors capture texts that may sound strange but grammatically correct.

This category is subdivided in the following subcategories: (1) orthographic, that captures the wrong use of accentuation or misspellings. Moreover, we include in this category other writing mistakes, such as starting sentences in lowercase; (2) stylistics, that captures some bad habits in writing, such as starting sentences with cardinal numbers or repeating multiple sentences with the same word; and (3) performance errors, that captures the wrong use of punctuation symbols.

This linguistic category contains 15 linguistic features. Table 2.3 shows some examples regarding correction and style, including stylistic and performance errors.

TABLE 2.3: Some features and examples of Correction and Style category

Features and **Examples**
`orthographic-sentences-starting-in-lowercase` el perro de mi amigo
`orthographics-misspelled-words` denonio, danto, varcelona
`orthographics-misspelled-accents-words` camion
`sentences-starting-with-numbers` 3 personas vinieron a la tienda
`sentences-starting-with-the-same-word` Creo que no es cierto. Creo, sin embargo, que no lo hizo a drede
`performance-duplicated-words` asique, enserio, portanto, sobretodo, sin fin, hechar, duplex
`performance-redundant-expressions` ambos dos, colaboración mutua, conclusión final, opinión personal

- **Semantics (SEM)**. Semantics is related to the intented meaning. In linguistic, semantics can be arranged at word, sentence or discourse level. UMUTextStats captures four linguistic features: (1) onomatopoeia, that are words created from the sound associated with what is named; (2) euphemisms, that are mild expressions that replace other words that can be considered too rude; (3) dysphemisms, that are derogatory expressions used instead of a pleasant one; and (4) synecdoches, that are figures of speech in which we use only a part but with the intention of representing the whole concept. Table 2.4 includes some examples of these linguistic features.

- **Pragmatics (PRA)**. Pragmatics is about how language is used and the context within. In this category we capture several features related to the usage of figurative language. These features include hyperboles, idiomatic

TABLE 2.4: Some features and examples of Semantics category

Features	Examples
onomatopoeia	boo+m, cataplún, crack, glub
euphemisms	ataque preventivo, capital humano, personas? de color
dysfemisms	estirar la pata, tarugos?, caja tonta
synecdocs	cabezas? de ganado, traer el pan

expressions, understatements, verbal irony, metaphors and similes, and some rhetorical questions. UMUTextStats has also some features for capturing discourse markers. These markers are used for structuring the conversation. UMUTextStats distinguishes among adders, reformers, argumentative clauses, or conversational-bookmarks, just to name a few. Besides, the tool captures many typical courtesy forms used in greetings or condolences.

This linguistic category has a total of 32 linguistic features. Some examples of features related to Pragmatics are depicted in Table 2.5.

TABLE 2.5: Some features and examples of Pragmatics category

Features	Examples
Figurative Language	
hyperboles	millones de veces, montañas de trabajo
idiomatic-expressions	abogado del diablo, dedo en la llaga
rhetorical-questions	¿Qué queréis que os diga?
verbal-irony	estar como una regadera
understatements	menos mal, tampoco es tan
metaphors	cabello de oro, corazón de cristal
similes	tus ojos son como estrellas
Discourse markers	
structuring-commenters	pues bien, dicho esto
structuring-order	en primer lugar, por un lado
connectors-additive	es más, por cierto
connectors-consecutive	por tanto, entonces, de ahí
connectors-reformers-corrective	mejor dicho, más bien
connectors-reformers-distance	en todo caso, de todos modos
Courtesy forms	
greetings	bienvenidas, mucho gusto, encantado
requirements	podría, le importaría, con permiso
condolences	sentido pésame, sentir la pérdida

- **Stylometry (STY)**. This category contains linguistic features concerning the linguistic style in written communication. Specifically, we include measures of the length of the text, diverse formulas for the Type-Token Ratio (TTR) (standard and normalised) [17].

In addition, there are features that measure the number (or percentage) of words, syllables, sentences, or uppercase letter. We also include some readability formulas and a rich variety of punctuation symbols.

There are 87 linguistic features within this linguistic category. The reader can find some examples of stylometry in Table 2.6.

TABLE 2.6: Some features and examples of Stylometry category

Features	**Examples**
Word statistics	
length	Hola mundo (10)
uppercase	hola MUNDO (50%)
expressions-within-parenthesis	2
words-longer-6tr	El caballo triste 33.3%
Sentence statistics	
count	Hola. ¿Cómo estás? (2)
interrogative-percentage-emphasis	¡¡Hola!! (100%)
Symbols and punctuation	
punctuation-symbols-currencies	€, $
punctuation-symbols-pipe	|

- **Lexis (LEX)**. This category includes dictionaries of words and expressions concerning specific topics. There are features for general domains with the intention of capturing the topic of the message. So, we have topics related to jobs, animals, crime, wealth, achievement or risk.

  This linguistic category is subdivided into 48 linguistic features. Table 2.7 contains some examples of features related to lexis.

TABLE 2.7: Some features and examples of Lexis category

Features	**Examples**
locations	Paris, Madrid, España
organisations	OMS, Indetex, Microsoft
animals	perro, gatos, conejos
weapons	cuchillo, fusil, escopeta
jobs	abogado, electricista
body	manos, pies, ombligo
death	muerte, fallecimiento, cementario
home	hogar, pasillo, rellano

- **Psycho linguistic processes (PLP)**. With this category, we intend to capture features from a cognitive point of view; that is, concerning language comprehension, production, and acquisition. We focus on features that capture positive and negative emotions and attitudes as well as other emotions.

  This linguistic category is subdivided into 11 linguistic features. Next, we include Table 2.8 with some instances concerning Psycho linguistic processes.

TABLE 2.8: Some features and examples of Psycho-linguistic processes category

Features	Examples
positive	feliz, alegre, genial
pleasure	gusto, bailar, deporte
negative	triste, enfadado, furioso
anger	amenazas, armas, asco

- **Register (REG)**. This category captures features related to register, which indicate how an speaker or a writer uses the language under different circumstances. We capture features related to offensive and informal.

  There are 13 different linguistic features within this linguistic category. Refer to Table 2.9 for some examples of this linguistic category.

TABLE 2.9: Some features and examples of Register category

Features	Examples
offensive-speech	bocazas, tonto, gilipollas
colloquialisms	acabose, anda ya, paća, pan comido
non-fluent	hmm, la movida esa, ajá
cultisms	fagocitar, cefalea, astronomía
latinisms	climax, status, a priori

- **Social media (SOC)**. As the majority of the experiments performed are from documents extracted from social networks, we decided to create an extra category to capture the extent in which users of social networks communicate. Specifically, we capture the percentage of hashtags, mentions, and hyperlinks.

  Social media category is subdivided into 9 linguistic features. In Table 2.10, some examples of this linguistic category are listed.

TABLE 2.10: Some features and examples of Social media category

Features	Examples
hashtags	#pln #deeplearning
mentions	@user1, @user2
urls	`http://www.example.com`
jargon	trolls, hashtags, follow, tuitstars

Besides, in Appendix A the complete list and taxonomy of the developed linguistic features organised by categories are listed, including a description of each one. It is worth mentioning that there is a total of 394 linguistic features. However, some of these linguistic features are only used for categorising and do not represent any linguistic phenomena. In other cases, some linguistic features have repeated values, as happens with `phonetics` and `phonetics-expressive-lengthening` because it is the only linguistic feature of `phonetics`. After discarding these non-informative

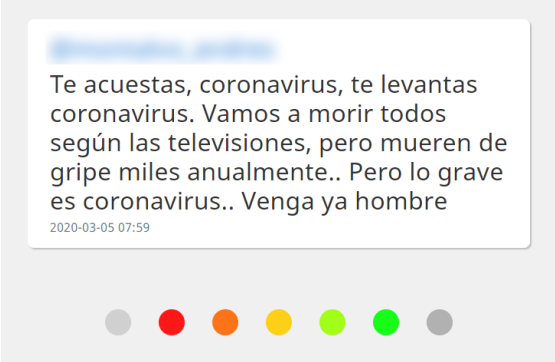

FIGURE 2.1: UMUCorpusClassifier. Screenshot of an annotator's view

features, we report that the real number of linguistic features reported by UMUTextStats is 365.

2.4 UMUCorpusClassifier

In order to evaluate UMUTextStats, we developed another linguistic tool, UMUCorpusClassifier, whose objective is the compilation and annotation of linguistic corpora [47]. This tool was developed because the compilation of annotated corpora is a time-consuming task. Besides, the quality of the corpus is heavily influenced by disagreements between annotators. Therefore, the lack of supervision of the annotation process can lead to poor quality corpora.

UMUCorpusClassifier is mainly focused on Twitter, a micro blogging social network that is popular in order to conduct document classification tasks [2]. This tool uses the Twitter search API in order to extract tweets based on a search string and, optionally, a geographic location.

Once the documents are compiled, they can be classified using two strategies: distant supervision and manual labelling. On the one hand, distant supervision allows to define rules to classify the documents based on certain conditions. We apply this strategy, for instance, to label automatically as satirical those tweets written by satirical newspaper (assuming that all those tweets are satirical). On the other hand, for performing a manual labelling stage, UMUCorpusClassifier allows to coordinate groups of annotators. Thus, the quality of the resulting corpora varies based on the number of annotators who classify the same tweet. UMUCorpusClassifier shows different metrics that allows to analyse the performance of the inter-annotator agreement, such as Krippendorff's alpha [56]. The allowed labels are highly configurable by corpora. It is worth noting that this tool can be easily extended to use other data sources.

The UMUCorpusClassifier tool allows to export the results in multiple formats. Besides, it can generate cleaned versions of the texts. For producing cleaned

versions of the text, the users may select to (1) remove blank lines, (2) strip HTML tags, (3) remove URLs, mentions and emojis, (4) remove letter elongations, (5) convert the texts to their lowercase form, or (6) fix misspellings automatically. In a nutshell, this process involves the analysis of each word in isolation. Next, it replaces the incorrect word with the best suggestion but only if a text similarity measure is higher that certain threshold.

2.5 Experimental results

In this section, the domains in which the linguistic features from the UMUTextStats tool have been evaluated are described. These features have been evaluated separately and combined with embedding based features into traditional machine-learning models and modern deep-learning architectures.

This section is divided into subsections. Each subsection is about a different domain and contains the published articles related to this doctoral thesis. Besides, we include other results of the same domain presented in other articles and the participation in international workshops. A summary of all the publications related to this doctoral thesis can be found in Chapter 4.

2.5.1 Aspect-based Sentiment Analysis

This section describes the validation of the linguistic features to the Sentiment Analysis task. Specifically, we evaluate the linguistic features in an aspect-based sentiment analysis study in the health domain. This study is concerning infodemiology, which is focused on the usage of information available on the Internet in order to improve health services [25]. Next, we evaluate other tasks regarding sentiment analysis concerning the financial domain or the participation in tasks related to emotion and sentiment analysis.

Infodemiology

To evaluate the infodemiology domain, we compiled a dataset from Twitter with short texts related to different infectious diseases. The tweets were compiled from Ecuador from keywords such as *Zika* or *Chikungunya*. The dataset was compiled and annotated using the UMUCorpusClassifier tool. The classification stage was performed by 20 students from the University of Guayaquil who performed 51 127 manual annotations. Each document was labelled an average of 6.0216 times, achieving a inter-coder reliability of 0.6864 based on Krippendorff's Alpha. We used this information to identify and discard those documents with less consensus. The final dataset contains 10 843 positive, 10 843 negative, and 7 659 neutral tweets.

The developed dataset is described in Table 2.11. This dataset is the one used for conducting the aspect-based sentiment analysis concerning infodemiology [33].

This dataset is available to download[6]. The corpus and more details regarding its compilation and annotation can be found at [33].

Name	Labels
positive	10 843
neutral	10 843
negative	7 659
Total	29 345

TABLE 2.11: Figures of the dataset developed for aspect-based Sentiment Analysis focused on Infodemiology [33]

Once the dataset was compiled, we extracted the linguistic features and used them to perform a sentiment analysis with three levels of sentiments: negative, neutral, and positive. The results obtained are reported in Table 2.12, in which the accuracy of a ten-fold cross validation is reported. As it can be observed, out the features evaluated, LF achieved an accuracy of 55.3%. These results outperform the rest of the features, which includes non-contextual word embeddings trained with a convolutional or a recurrent neural network.

Our next step consisted in aspect-level part of the system. The aspects related to infodemiology were represented within an ontology. This ontology contains classes related to risks, symptoms, transmission methods or drugs among others. It is worth mentioning that this ontology was designed from scratch for this project, although it was based on standards and other ontologies such as Disease Ontology [89] and the Infectious Disease Ontology [20].

As we deal with short texts, we assume that one tweet contains only one sentiment. Accordingly, we ranked the relationship between the sentiment of the tweet with the ontology classes. Our main objective was to measure how much a concept influences others. For this, we used an extended version of the TF–IDF formula [82] (see Eq 2.1), in which each topic influences the TF value of the rest of topics of the ontology. The degree of this contribution is less important as more distant the concepts of the ontology are. For example, if we label a document as negative related to fever, this

[6] https://umuteam.inf.um.es/corpora/misogyny/zika-spanish-2020.rar

TABLE 2.12: Results of the aspect-based sentiment analysis concerning infodemiology. The results are ranked by accuracy

Model	**k1**	**k2**	**k3**	**k4**	**k5**	**k6**	**k7**	**k8**	**k9**	**k10**	**AVG**
LF	57.1	55.4	55.6	55.1	51.9	56.1	52.3	56.7	57.0	55.9	**55.3**
LSTM	52.9	56.4	33.3	33.3	46.8	49.5	49.3	47.1	50.8	48.6	**46.8**
LSTM+LF	53.4	63.6	55.7	46.4	44.6	47.7	48.7	52.5	49.7	47.6	**51.0**
BiLSTM	33.2	51.7	33.4	52.2	52.9	33.2	33.2	51.8	53.6	33.5	**42.9**
BiLSTM+LF	52.3	52.1	56.5	52.4	53.9	57.0	56.6	52.3	51.9	56.7	**54.2**
CNN	51.2	56.0	49.9	45.6	45.6	48.7	50.6	47.7	49.1	48.1	**49.3**
CNN+LF	53.1	53.6	48.3	46.4	46.8	48.2	51.3	44.6	50.0	48.4	**49.1**

negative sentiments greatly affects the concept of fever, since it appears explicitly in the text. It also influences those diseases whose symptom is fever, but in a lesser degree.

In order to get insights regarding the relationship of the linguistic features and the sentiments, we calculated the Information Gain [84]. Figure 2.2 depicts the results achieved. The results are ordered by relevance but normalised. This means that each bar represents the percentage to which each feature contributes to the positive, neutral or negative labels. These findings allowed us to verify that numerals are correlated to negative tweets, as numerals are used to report news related to official data of deceased or infected. We also observed that the usage of colloquialism is more related to positive and neutral tweets than negative tweets.

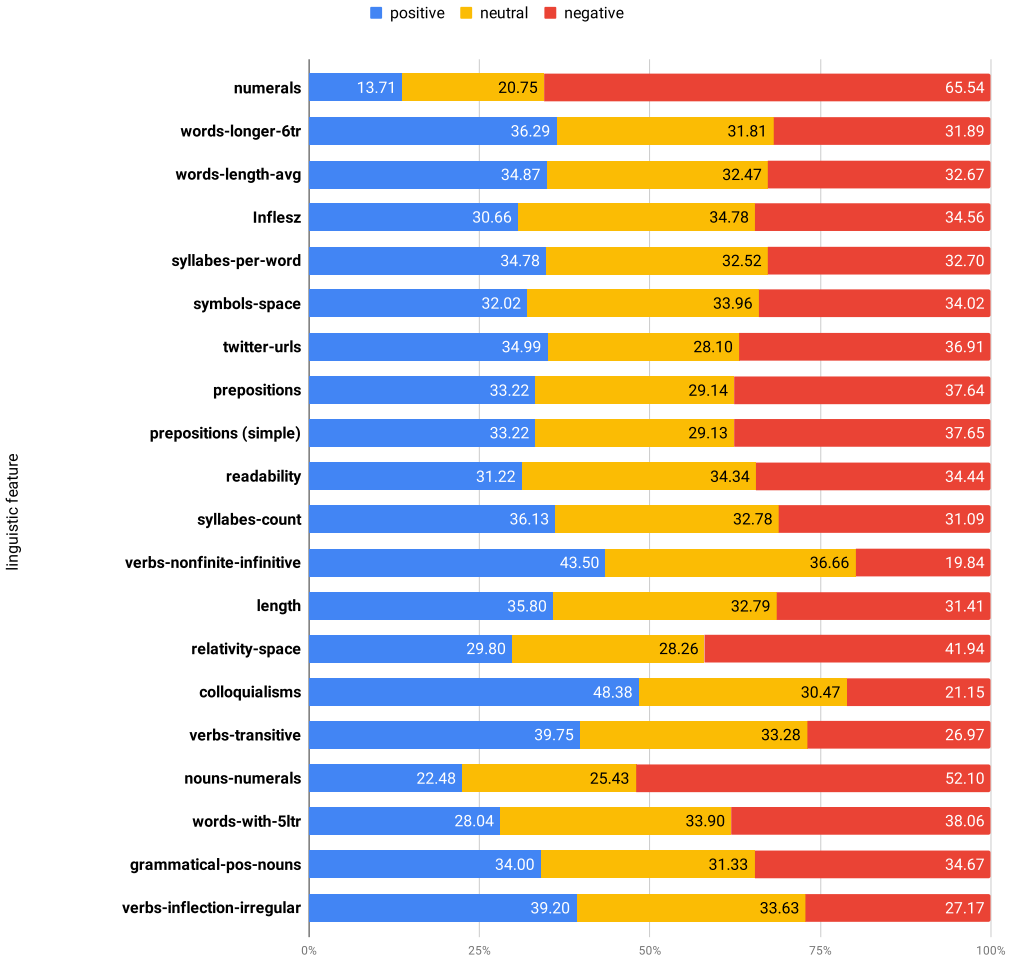

FIGURE 2.2: Information gain, grouped by sentiment and linguistic feature[33]

TABLE 2.13: EmoEvalES 2021: Official results of the task, ranked by accuracy

Rank	Team/User	Accuracy	M-precision	M-recall	M-F1 score
1	daveni	72.7657	70.9411	72.7657	71.7028
2	fyinh	72.2222	70.4695	72.2222	71.1373
3	HongxinLuo	71.2560	70.4496	71.2560	70.5432
4	JorgeFlores	70.2899	69.2397	70.2899	69.6675
5	hahalk	69.2029	67.9620	69.2029	66.3740
6	**UMUTeam**	**68.5990**	**67.2546**	**68.5990**	**66.8407**
7	ffm	68.4179	68.2765	68.4179	68.2487
8	fazlfrs	68.2367	66.4868	68.2367	66.8757
9	luischir	67.8140	65.8314	67.8140	65.7367
10	vitiugin	67.5725	65.7681	67.5725	66.1427
11	job80	66.8478	65.2840	66.8478	64.6085
12	aridearriba	65.2778	60.0479	65.2778	62.2223
13	Timen	61.7754	59.7877	61.7754	60.0217
14	QuSwe1d0n	53.6836	65.3707	53.6836	55.7007
15	qu	44.9879	61.8833	44.9879	44.6947

Other contributions regarding Sentiment Analysis and Emotion Analysis

The linguistic features of UMUTextStats were evaluated in other Sentiment Analysis works, but not from an aspect-based perspective. Our first evaluation was conducted in the shared-task TASS [48], proposed in IberLEF 2020. Specifically, two challenges were proposed: (1) an automatic classification problem based on three levels (positive, neutral, and negative) of tweets written in Spanish from Spain and Latin America; and (2) a multi-classification task based on determining six basic emotions from Ekman [24]. During TASS, we sent three runs that combined the linguistic features with pre-trained sentence and word-embeddings with convolutional neural networks. Although our results for the first challenge were limited, we achieved the best precision in the second challenge. The reader can find more details about our participation in [31].

The evaluation of the linguistic features in emotion analysis was also conducted in 2021 in the EmoEvalEs shared task [6], in which we ranked in the sixth position (see Table 2.13), achieving an accuracy of 68.5990% (4.1667% below the best result).

We also evaluated the linguistic features applied to Sentiment Analysis in the financial domain. This domain is particularly hard, as documents concerning finances usually contain expressions whose meaning depends heavily on the context, hindering Sentiment Analysis. For this, we compiled a dataset from different news sites and experts in economy using the UMUCorpusClassifier tool. Preliminary versions of this dataset were used to evaluate non-contextual Spanish word and sentence embeddings from Spanish pre-trained models [44, 32]. However, we continued with the compilation and annotation of financial tweets and compiled a final dataset with 15 915 tweets labelled as positive, negative and

neutral. As a second step, we evaluated several contextual word and sentence embeddings based on Transformers. These results have been sent to a high-impact scientific journal and we are waiting for the first review.

2.5.2 Hate-speech and misogyny detection

In this section we describe the contribution of the linguistic features to hate-speech detection and other related tasks such as misogyny identification. Hate-speech detection has become in the last years a trend in workshops related to NLP. Due to the large number of posts published daily in social networks and the inability to review them by hand, automatic hate-speech detection is a need tool in order to keep social environments safe from misogynistic, xenophobic, and homophobic people that intimidate people because of their gender, ethnicity or sexual orientation.

Next, we describe the compilation process and validation of the Spanish MisoCorpus 2020, a dataset concerning misogyny identification in Spain, which is the main contribution of this doctoral thesis concerning the hate-speech domain.

Misogyny detection

We have evaluated our methods in different studies regarding misogyny detection. Our first contribution in this field was the compilation, annotation, and evaluation of the Spanish MisoCorpus 2020 [32]. We released this dataset as a whole and divided into three splits. The first split, VARW, is concerning violence against relevant women, focused on aggressive messages on Twitter to women who have gained social relevance. The second split, SELA, is focused on discerning misogynistic messages from Spain and Latin America. The third split, DDSS, contains general traits related to misogyny, namely discredit, dominance, sexual harassment and stereotype. This dataset was also compiled using the UMUCorpusClassifier tool and was manually annotated by three members of our research group. The dataset is balanced and it contains 3 841 misogynous documents.

Table 2.14 depicts the statistics of the Spanish MisoCorpus 2020 [42] and its three splits, namely VARW, SELA, and DDSS[7].

Name	Misogyny	Not misogyny
VARW	2094	2094
SELA	2081	2081
DDSS	1665	1665
MisoCorpus-2020	3841	3841

TABLE 2.14: Spanish MisoCorpus 2020

[7] https://umuteam.inf.um.es/corpora/misogyny/misocorpus-spanish-2020.rar

TABLE 2.15: Resume of the results achieved with the Spanish MisoCorpus 2020, organised by machine learning classifier, model and split.

Classifier	Model	VARW	SELA	DDSS	SMC-2020
RF	BoW	78.930	76.967	74.734	76.215
	SE	**82.092**	**81.307**	**79.063**	77.232
	LF	81.112	81.740	77.613	79.237
	SE+LF	82.092	81.307	78.912	**79.302**
SMO	BoW	78.524	76.918	74.003	73.798
	SE	**84.886**	82.100	**81.360**	81.020
	LF	82.403	80.057	77.976	78.938
	SE+LF	84.886	**85.175**	81.208	**85.175**
LSVM	BoW	80.053	78.476	77.698	77.060
	SE	**84.480**	**81.859**	**81.148**	80.825
	LF	82.283	81.115	79.245	79.263
	SE+LF	84.480	83.734	80.755	**82.882**

This work was one of the first ones to evaluate the linguistic features. Thus, certain techniques were not yet mastered and models based on deep-learning were not used. Instead, we based our study in traditional machine learning approaches. Specifically, the LF was combined in isolation or combined with sentence word embeddings from fastText [50].

The best performance of our systems was achieved with Support Vector Machine (SVM), achieving an accuracy of 85.175%. In Table 2.15 we can observe the accuracy for each feature set evaluated with the SVM. This comparison involved a baseline model based on BoW, the average word embeddings from fastText (AWE), the linguistic features in isolation (LF), and the combination of linguistic features and the average of words embeddings (AWE+LF). We can observe that the combination of linguistic features and the average of word embeddings outperformed the rest of the feature sets which supports our first hypothesis regarding the improvement of the results for text classification tasks.

In this work we also used the linguistic features for the interpretability of the results. Specifically, in Figure 2.3 we include the normalised values of Information Gain for the 20 best linguistic features for the MisoCorpus-2020. So, we can observe whether these values correspond to the *misogyny* or *not-misogyny* labels. We observed that features related to register and, specifically, offensive language, have a strong correlation for misogyny detection. We also found a strong correlation between the grammatical gender and misogyny identification. This is important because, in Spanish, some words can be interpretated differently according to their gender. This happens with male and female names of animals that denote different traits of a person. For example, a male fox (*zorro*) describes a clever person whereas a female box (*zorra*) denotes a female prostitute. We also observed a strong correlation with correction and style features, such as the percentage of misspelled

words. This fact suggests that the correct usage of language is a relevant factor to consider in misogyny identification. Moreover, *informal-speech-collocations*, which are redundant expressions, suggest that users do not take enough time to consider and prepare their arguments. Therefore, this feature (their impulsive speaking) contributes to the detection of misogynous statements. In addition, we also achieve relevant findings related to PoS features, such as the percentage of qualifying adjectives and adverbs.

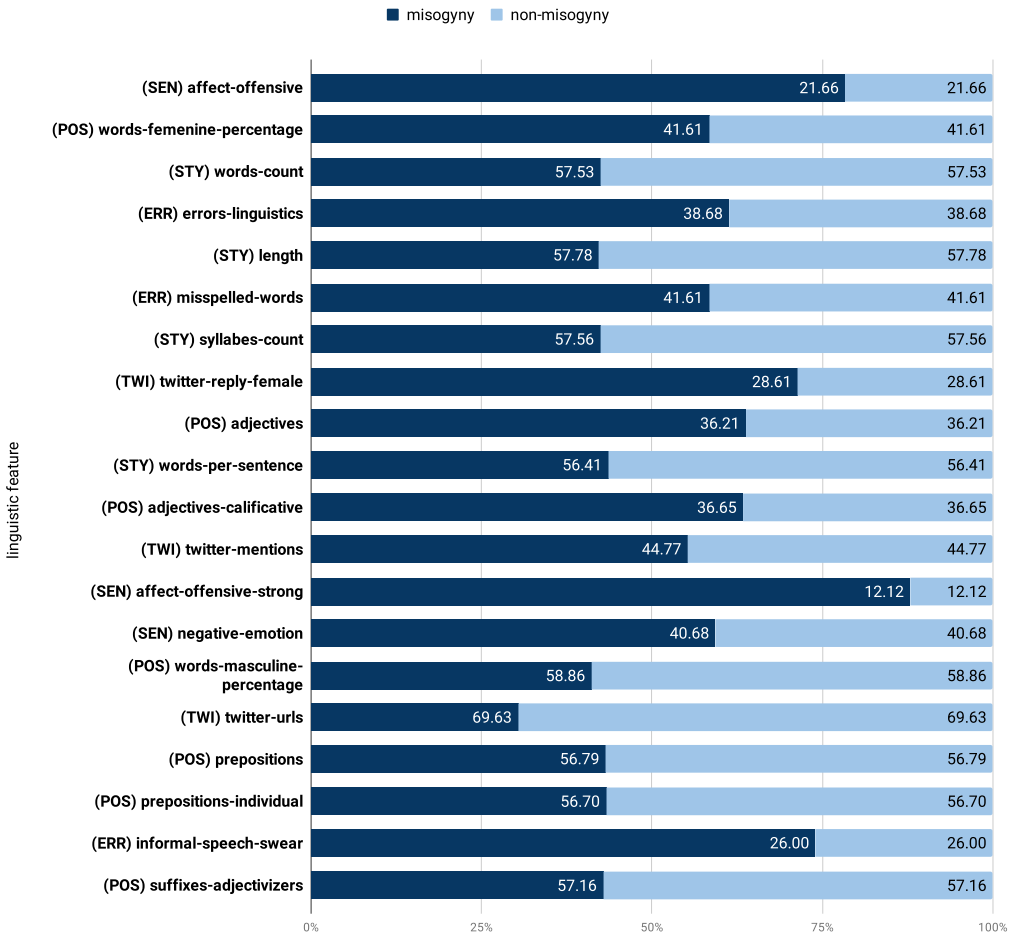

FIGURE 2.3: Information gain, grouped by sentiment and linguistic feature related to misogyny identification using the Spanish MisoCorpus 2020 [32]

In this study, we also evaluated our methods with two external datasets. First, we evaluated the AMI'2018 dataset [28]. With this dataset, we outperformed the best results by combining the LF and SE the word embeddings with a linear SVM. Second, we evaluated HatEval'2019 [11]. With this dataset, our proposal outperformed the baseline proposed as well as the best results of the participants of the shared task with the combination of the average of word embeddings with the linguistic features with an accuracy of 75.4505%.

Other contributions regarding Hate-speech

We also have participated in the detection of online sexism within the EXIST shared task from IberLEF 2021 [83]. This shared task proposed two challenges: a binary sexism classification and a multiclass problem for sexism categorisation. This dataset contained documents compiled from Twitter and gabs in two languages: Spanish and English. We achieved position 16, with an accuracy of 75.14%, improving all baselines proposed and only nearly 2-3% below the best result.

Besides, we have evaluated together the LF with negation features along with Spanish pre-trained contextual and non-contextual embeddings for detecting hate-speech in Spanish. These features were evaluated with several datasets concerning hate-speech and misogyny in Spanish. These datasets are the Spanish MisoCorpus 2020, the Spanish partitions of the datasets published in the AMI 2018 and HatEval 2019 shared tasks, and HaterNET [71]. The results and conclusions of this research are presented at [43].

The methods described in this paper were also employed in the shared tasks, regarding the identification of offensive language, in the domain of hate-speech detection. These shared tasks were MEX-A3T 2020 [5] and MeOffendEs 2021 [75].

The first shared-task, MEX-A3T, consists in the identification of aggressiveness in tweets written in Mexican Spanish. Our proposal for solving this task was grounded on the combination of the LF with Spanish pre-trained word embeddings from GloVe, fastText and word2vec, that is, the non-contextual word embeddings. We achieved limited results in this task, as none of or runs outperformed the baseline proposed.

The second shared-task, MeOffendEs 2021, was focused on the identification and categorisation of offensiveness. MeOffendEs consisted in two subtasks. On the one hand, a multi-classification for the European Spanish subtasks, discerning whether the offensive texts whose target is a person or group, just use inadequate language, but not necessarily offensive, or nothing of the above. On the other hand, the Mexican Spanish dataset consisted into a binary classification problem. Both challenges included a variant in which contextual features from the documents could be considered. In this shared-task, we could do a collaboration with the Universidad de Jaén. Specifically, we combined our methods with fine-grained negation features [54].

In order to combine all these features, we relied on ensemble learning. Specifically, we evaluated the following strategies: based on the mode of the predictions, ensembles based on averaging the predictions of each neural network, ensembles based on the highest probability, and ensembles based on training regression machine learning model from the probabilities of the training split. We observed that the ensembles based on linear regression provided the best results whereas the

ones based on the highest probability provided the best precision over the offensive class.

We ranked in the 2nd place in subtask 1 (with a F1-score of 87.8289%), 1st in subtask 2 (F1-score 87.8289%), 5th in subtask 3 (F1-score of 67.0588%), and 1st in subtask 4 (F1-score of 66.9449%). However, there were less participants in the subtasks that included the contextual features. Regarding the interpretability of the models, we observed in the Spanish dataset that negative psycho-linguistic processes were strong features to discern from non-offensive documents from the others, but that they were not good indicators to discern among if the target is a person, a group or simply the use of inadequate language.

It is worth mentioning that we evaluated a subset of the linguistic features in the HASOC shared task [69], concerning offensive language in English, Hindi and Marathi. For this, we combined the features from the stylometry category with BERT. This shared task proposed two challenges, but we only participated in the first one, that consisted in a binary classification problem to spot social posts with hateful or offensive content, and a multi-classification challenge problem to discriminate between hate, profane and offensive traits. For the binary classification problem, we reached a macro F1-score of 80.13% in English, a 75.20% in Hindi, and a 84.23% for Marathi. In case of English, we achieved our best result using plain BERT. For the Hindi and Marathi, we achieved our best results using ensemble learning. However, for the multi-classification challenge, we got the best results with ensembles, achieving a macro F1-score of 62.89% for English, and a 51.67% for Hindi.

2.5.3 Figurative language. Satire, Sarcasm, and Humor detection

One particular challenge concerning NLP in general is the figurative language, in which words deviate from their conventional meaning. Figurative language is present in literary genres such as satire, or in sarcastic statements, very popular in social networks. Due to its relevance, figurative language has been analysed carefully. Sarcasm, among other forms of figurative speech, such as irony, and literary forms, such as satire, has been explored in [73], in which the authors explore the most discriminant features for satire and irony detection.

Next, we describe our contributions in satire, sarcasm and humour detection.

Satire identification from real-news

Satire is a literary genre. We should not think that satire is only for our entertainment. In contrast, satire is a powerful tool that allows citizens to overcome their weaknesses. However, satirical news is often wrongly classified by fake-news detectors as fake news. Accordingly, we evaluate how effective the linguistic

features extracted from UMUTextStats are in to distinguish between satirical news and real news.

We follow the methodology exposed in [72] and [9] to compile the Spanish SatiCorpus 2021. This dataset is balanced and contains news headlines from Twitter compiled with the UMUCorpusClassifier tool. The accounts were selected from different Spanish spoken countries. Moreover, we decided to enlarge this dataset including tweets from Twitter accounts used for impersonate and satirise real relevant people. This dataset was automatically annotated, based on the idea that all tweets from satirical news media are satiric. The Spanish SatiCorpus 2021 contains 18 207 satiric and 18 207 non-satiric tweets and contains tweets between March, 2018 to June, 2021.

Table 2.16 is the corpora used for conducting the satire classification task described at [37]. This dataset is available at https://umuteam.inf.um.es/corpora/satire/spanish-saticorpus-2021.zip.

Label	Train	Development	Test	Total
satire	10 923	3 642	3 642	18 207
non-satire	10 923	3 642	3 642	18 207
total	21 846	7 284	7 284	36 414

TABLE 2.16: Corpus distribution per label and split for the Spanish SatiCorpus 2021

The next step aimed at evaluating the LF with the Spanish SatiCorpus 2021. We evaluated the LF separately and combined with different types of features using different strategies. Our best result is achieved with a combination of the LF and BERT with an accuracy of 97.405%. We include in Table 2.17 some of the results achieved for linguistic features in isolation or combined using knowledge integration and ensemble learning.

Figure 2.4 contains the Information Gain of the linguistic features using the Mutual Information measure. From the correction and style category, it can be observed that the number of orthographic errors is more common in non satirical documents than in satirical documents. In contrast, the number of hashtags more commonly appears in non-satirical documents. Regarding morphological features, the use of pronouns and nouns is good for discerning between satirical and non-satirical documents, being the pronouns more frequently found in satirical documents whereas nouns are more common in non-satirical documents.

Humor identification. HaHa and Hahackathon 2021

We also evaluated the linguistic features in two shared tasks regarding the identification, categorisation, and evaluation of humor: Hahackathon [65], proposed at SemEval 2021 and focused on English, and HaHa 2021 [18], proposed at IberLEF 2021 and focused on Spanish. Both challenges were divided into four

TABLE 2.17: Precision, recall, F1-score of satiric and non-satiric labels by feature set separately. Macro-averaged precision, recall, F1-score, and accuracy of the overall result [37]

Feature set	Precision	Recall	F1-score	Accuracy
(LF) Linguistic features				
non-satire	86.240	83.635	84.918	-
satire	84.115	86.656	85.367	-
macro-avg	85.178	85.146	85.142	85.146
Knowledge integration of LF + SE + BF				
non-satire	97.782	96.842	97.310	-
satire	96.872	97.803	97.336	-
macro-avg	97.327	97.323	97.323	97.323
Knowledge integration of LF, SE, WE and BF)				
non-satire	97.281	97.254	97.268	-
satire	97.255	97.282	97.268	-
macro-avg	97.268	97.268	97.268	97.268
Ensemble learning: Highest probability				
non-satire	99.752	66.310	79.664	-
satire	74.769	99.835	85.503	-
macro-avg	87.260	83.072	82.583	83.072
Ensemble learning: Weighted mode				
non-satire	95.623	95.387	95.505	-
satire	95.399	95.634	95.516	-
macro-avg	95.511	95.511	95.511	95.511
Ensemble learning: Probability average				
non-satire	93.332	94.152	93.740	-
satire	94.100	93.273	93.685	-
macro-avg	93.716	93.712	93.712	93.712

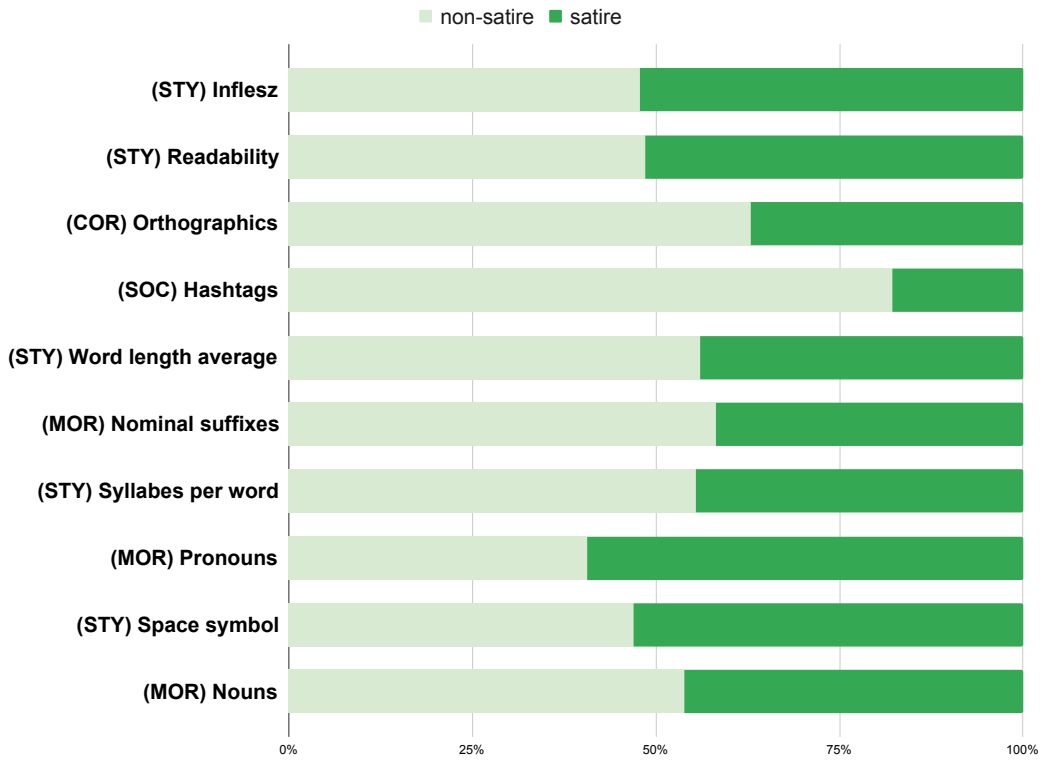

FIGURE 2.4: Information gain for satiric and non-satiric documents [37]

minor challenges. The HaHackathon shared task focused on determining when a text is funny, how funny it is (regression), and if it contains some offensive content and, if so, how much. HaHa 2021 shared with HaHackathon the first two challenges, but replaces the controversial humor with two new challenges focused on determining what are the mechanisms to make a text funny and what are the targets of the joke.

Our participation in the Hahackathon 2021 shared task is described at [41]. In a nutshell, we reached position 45 in subtask 1a (F1-score of 91.60%). Position 47 in subtask 1b (RMSE of 0.8847). Position 14 in subtask 1c (F1-score of 57.22%), and position 46 for subtask 2a (RMSE of 0.8740). It is worth noting that as this shared task contains only English documents, we only use the subset of the LF based on corpus statistics and stylometry. Our participation in the Haha 2021 shared task is described at [39]. We reached the 1st position in Funniness Score Prediction. Position 8 in humor classification. Position 7 in the humour mechanism detection, and position 3 in the humour target classification. See Table 2.18 for a comparison with the rest of the participants.

In addition, we have participated in the iSarcasm 2022 shared task, proposed in SemEval 2022 and that contains documents in English and Arabic. Our results in this shared-task were limited for English (reaching position 41) but better for

TABLE 2.18: Official results and ranking of the HAHA'2021 task for each subtask, ranked, respectively by F1 score for the humorous category (task 1), RMSE (Task 2), and macro F1-score (Task 3 and 4)

Team / User	Subtask 1	Subtask 2	Subtask 3	Subtask 4
Jocoso	88.50 (1)	0.6296 (3)	0.2916 (2)	0.3578 (2)
icc	87.16 (2)	0.6853 (9)	0.2522 (3)	0.3110 (4)
kuiyongyi	87.00 (3)	0.6797 (8)	0.2187 (5)	0.2836 (6)
ColBERT	86.96 (4)	0.6246 (2)	0.2060 (7)	0.3099 (5)
noda risa	86.54 (5)	-	-	-
BERT4EVER	86.45 (6)	0.6587 (4)	0.3396 (1)	0.4228 (1)
Mjason	85.83 (7)	1.1975 (11)	-	-
UMUTeam	**85.44 (8)**	**0.6226 (1)**	**0.2087 (6)**	**0.3225 (3)**
skblaz	81.56 (9)	0.6668 (6)	0.2355 (4)	0.2295 (7)
humBERTor	81.15 (10)	-	-	-
RoBERToCarlos	79.61 (11)	0.8602 (10)	0.0128 (10)	0.0000 (9)
lunna	76.93 (12)	-	0.0404 (9)	-
N&&N	76.93 (12)	-	0.0404 (9)	-
ayushnanda14	76.79 (13)	0.6639 (5)	-	-
Noor	76.03 (14)	-	0.0404 (9)	-
KdeHumor	74.41 (15)	1.5164 (12)	-	-
baseline	66.19 (16)	0.6704 (7)	0.1001 (8)	0.0527 (8)

Arabic (reaching position 22). Our main limitation in this task consisted in that our proposals did not tackle the problem with the dataset imbalance, achieving limited macro F1-score for the sarcastic label.

2.5.4 Author analysis

Other research field concerning NLP in which the linguistic features from UMUTextStats were evaluated is author analysis. This research field aims to retrieve information from people based on their writings [22]. The applications of this research field are diverse. On the one hand, it can be used as linguistic evidence in forensic linguistics, as it can help to unmask who is the author of an anonymous threatening message. It can also be used in plagiarism detection to discern about who is the real author of a document, and even can help to determine whether a suicide note is real or not [4]. Author analysis can be categorised into three subtasks [55]. The first one is authorship attribution, focused on determining who is the author of a certain work. The second one is authorship verification, focused on determining if one specific author is the one who wrote certain document, by examining samples of other writings. The third one is author profiling, focused on the identification of certain traits of the authors. These traits are related to their age, their gender, and even their educational level, among others.

Next, we describe the validation of the UMUTextStats tool for conducting two author analysis tasks, concerning author profiling and author attribution.

Author profiling and author attribution in Politics

We have explored the reliability of the linguistic features in two experiments regarding author analysis: authorship attribution and author profiling. For this, we relied on UMUCorpusClassifier to compile samples of writing of Spanish politicians during 2020. We compiled almost 250k tweets from a total of 385 politicians. Next, we annotated each user with their gender, their year of birth, and their political spectrum on two axes (binary and multiclass). After discarding some of the tweets that were not written in Spanish or they were related to news sites, we selected the most representative tweets per user. For this, we grouped the tweets into twelve bins per user (one per month), and we organised the tweets in each bin according to a number of topics that appear in each document and their length. This way, we picked sequentially tweets for each bin until we got a minimum of 120 tweets per user. Finally, we anonymised the accounts of the politicians to hinder this task, and arranged the final dataset to do both author analysis tasks. In addition, in order to prevent bias, we compiled and extra dataset composed by tweets of journalists.

Table 2.19 depicts the distribution of the labels per user for conducting the author profiling task described in [34]. We include here the number of politicians per demographic (gender and age) and psychographic (binary and multiclass political spectrum). Besides, we include a list of journalists that were compiled in order to observe if our model was able to generalise the political spectrum in users that are not politicians. This dataset is also available[8].

The results achieved in the first task, author profiling, are summarised in Table 2.20, in which the results of two demographic and two psychographic traits are shown. In this case, we evaluated the LF along with sentence embeddings (SE), sentence BERT embeddings (SBE) -without fine tuning the model-, non-contextual pretrained word embeddings (PWE) and BETO, the Spanish BERT. We can observe that the LF achieved very promising results, outperforming BERT in some cases, such as gender prediction. Moreover, the combination of LF with any kind of embeddings usually results in better results than achieved with both features separately.

In a similar manner, we extracted the linguistic features and their correlation with the labels using the Information Gain for all demographic and psychographic traits (see Figure 2.5). It can be noticed that morphosyntax is the most relevant linguistic category for determining demographic traits. According to the age range trait, the number of personal pronouns is relevant and more common in younger politicians. Another relevant feature is related to colloquialisms, being more common in younger and older politicians but less frequent in middle-aged politicians. In addition, topics related to countries and languages that may indicate territorial policy issues appear also as relevant. Concerning gender, the percentage of verbs

[8] https://pln.inf.um.es/corpora/politics/policorpus-2020.rar

Trait	Class	Total	Train	Val	Test
		Politicians			
Gender	female	113	67	23	23
	male	156	99	29	28
Age	25-34	28	21	1	6
	35-49	126	80	23	23
	50-64	104	57	26	21
	over 65	11	8	2	1
Spectrum	left	146	88	31	27
(binary)	right	123	78	21	24
Spectrum	left	56	37	12	7
(multiclass)	m-left	90	51	19	20
	m-right	83	54	15	14
	right	39	23	6	10
		Journalists			
Spectrum	left	31	-	-	31
(binary)	right	20	-	-	20
Spectrum	left	20	-	-	20
(multiclass)	m-left	11	-	-	11
	m-right	13	-	-	13
	right	7	-	-	7

TABLE 2.19: Corpus distribution per label and split for the Spanish PoliCorpus 2020

TABLE 2.20: Results of the demographic and psychographic traits in an author profiling task

		Demographic		Psychographic	
Feature set	Architecture	$F1_{GENDER}$	$F1_{AGE\_RANGE}$	$F1_{PSBINARY}$	$F1_{PSMULTI}$
LF	MLP	68.4553	44.3834	86.2851	70.6511
SE	MLP	60.8751	34.3935	80.4072	54.4602
SBE	MLP	**70.3100**	39.3528	68.6516	55.1185
PWE	CNN	68.6275	41.9711	96.0632	81.9249
PWE	BiGRU	57.4160	27.1823	42.3630	21.4780
BETO	BERT	64.9020	48.3619	96.0784	80.4439
LF+SE	MLP	70.0002	40.1127	90.1961	76.5988
LF+SBE	MLP	66.3513	27.7193	82.2850	66.9801
SE+SBE	MLP	70.0002	35.6353	64.5689	62.5214
LF+PWE	CNN	64.7059	**49.3662**	90.1809	**82.5349**
LF+PWE	BiGRU	57.4160	30.2750	61.5170	25.8280
BETO+LF	BERT	65.8593	42.0046	**96.0784**	74.8282

and personal pronouns was relevant but we did not spot many differences among the gender trait. Concerning the psychographic traits, from a binary perspective, we observed that the topics were different. For example, politicians from the right-wing speak more about religion. In left-wing parties, on the other hand, we highlight the use of qualifying adjectives and features related to the spatial dimension. However, when looking at the political spectrum from a multiclass perspective, we observed a larger difference between the left and moderate left wing compared to the right and moderate right.

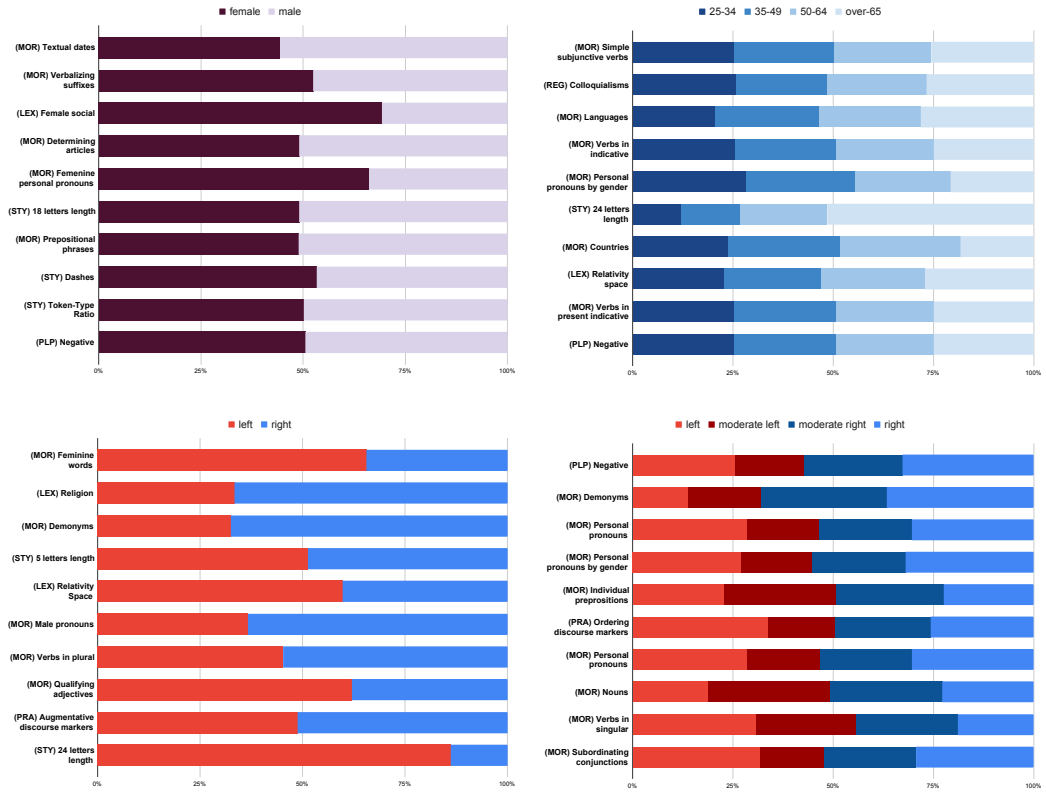

FIGURE 2.5: Information gain, grouped by demographic traits (top) showing the gender (left) and age range (right) and the psychographic traits (bottom) of political ideology in binary classification (left) and multiclass classification (right) [34]

Next, to conduct the authorship attribution task, we used the same tweets and politicians that are in the training set of the previous author profiling task. The results are depicted in Table 2.21.

It is worth mentioning that we have proposed a shared task in IberLEF 2022 entitled PoliticEs 2022[9] that consists into determining the psychographic and demographic traits of politicians and journalists. This dataset is an extension of the work described in [34].

[9] https://codalab.lisn.upsaclay.fr/competitions/1948

TABLE 2.21: Results of the authorship attribution task

Feature set	Architecture	$F1_{macro}$
NG	MLP	8.0939
LF	MLP	18.6417
SE	MLP	18.9682
SBE	MLP	20.8305
PWE	MLP	11.9486
PWE	CNN	8.1058
PWE	BiGRU	1.5146
BETO	BERT	27.2605
LF+SE	MLP	26.2711
LF+SBE	MLP	26.2557
SE+SBE	MLP	21.9318
LF+PWE	MLP	21.2380
LF+PWE	CNN	15.8582
LF+PWE	BiGRU	3.7828
BETO+LF	BERT	**29.3361**

Other contributions regarding Author Analysis

In addition, related to authorship verification, we participated in the AISOCO'2020 shared task from FIRE workshop [26] concerning authorship identification of source-code. The details regarding our participation can be found at [38]. In a nutshell, our participation dealt with character n-grams that we combined with author traits captured with UMUTextStats. We ranked the 6th position, with an accuracy of 91.16% in the official leader board. Moreover, we outperform baselines based on RoBERTa.

Chapter 3

Published articles

3.1 Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in Latin America

TABLE 3.1: Metadata of the first publication that compose this PhD Thesis

Key	Value
Title	Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in Latin America
Authors	José Antonio García-Díaz, Mar Cánovas-García, Rafael Valencia-García
Type	Journal
Journal	Future Generation Computer Systems
Impact Factor	7.187
Publisher	Elsevier
Pages	641–657
Volume	112
Year	2020
Month	November
ISSN	0167-739X
DOI	https://doi.org/10.1016/j.future.2020.06.019
URL	https://www.sciencedirect.com/science/article/pii/S0167739X2030892X
State	Published

3.1.1 Abstract

Infodemiology is the process of mining unstructured and textual data so as to provide public health officials and policymakers with valuable information regarding public health. The appearance of this new data source, which was previously unimaginable, has opened up a new way in which to improve public health systems, resulting in better communication policies and better detection

systems. However, the unstructured nature of the Internet, along with the complexity of the infectious disease domain, prevents the information extracted from being easily understood. Moreover, when dealing with languages other than English, for which some of the most common Natural Language Processing resources are not available, the correct exploitation of this data becomes even more difficult. We intend to fill these gaps proposing an ontology-driven aspect-based sentiment analysis with which to measure the general public's opinions as regards infectious diseases when expressed in Spanish by employing a case study of tweets concerning the Zika, Dengue and Chikungunya viruses in Latin America. Our proposal is based on two technologies. We first use ontologies in order to model the infectious disease domain with concepts such as risks, symptoms, transmission methods or drugs, among other concepts. We then measure the relationship between these concepts in order to determine the degree to which one concept influences other concepts. This new information is subsequently applied in order to build an aspect-based sentiment analysis model based on statistical and linguistic features. This is done by applying deep-learning models. Our proposal is available on a web platform, where users can see the sentiment for each concept at a glance and analyse how each concept influences the sentiment of the others.

3.1.2 Author's contribution

The PhD student, José Antonio García-Díaz, is the main author. He contributed in the conceptualisation with his thesis director, wrote the manuscript and actively contributed to its revision. At a technical level, the PhD student developed all the software for the execution of the experiments.

3.2 Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings

3.2.1 Abstract

Online social networks allow powerless people to gain enormous amounts of control over particular people's lives and profit from the anonymity or social distance that the Internet provides in order to harass other people. One of the most frequently targeted groups comprise women, as misogyny is, unfortunately, a reality in our society. However, although great efforts have recently been made to identify misogyny, it is still difficult to distinguish as it can sometimes be very subtle and deep, signifying that the use of statistical approaches is not sufficient. Moreover, as Spanish is spoken worldwide, context and cultural differences can complicate this identification. Our contribution to the detection of misogyny in Spanish is two-fold. On the one hand, we apply Sentiment Analysis and Social Computing technologies for detecting misogynous messages in Twitter. On the other, we have compiled the Spanish MisoCorpus-2020, a balanced corpus

TABLE 3.2: Details of the second publication that compose this PhD Thesis

Key	**Value**
Title	Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings
Authors	José Antonio García-Díaz, Mar Cánovas-García, Ricardo Colomo-Palacios, Rafael Valencia-García
Type	Journal
Journal	Future Generation Computer Systems
Impact Factor	7.187
Publisher	Elsevier
Pages	506–518
Volume	114
Year	2021
Month	January
ISSN	0167-739X
DOI	https://doi.org/10.1016/j.future.2020.08.032
URL	https://www.sciencedirect.com/science/article/abs/pii/S0167739X20301928
State	Published

regarding misogyny in Spanish, and classified it into three subsets concerning (1) violence towards relevant women, (2) messages harassing women in Spanish from Spain and Spanish from Latin America, and (3) general traits related to misogyny. Our proposal combines a classification based on average word embeddings and linguistic features in order to understand which linguistic phenomena principally contribute to the identification of misogyny. We have evaluated our proposal with three machine-learning classifiers, achieving the best accuracy of 85.175%. Finally the proposed approach is also validated with existing corpora for misogyny and aggressiveness detection such as AMI and HatEval obtaining good results.

3.2.2 Author's contribution

The PhD student, José Antonio García-Díaz, is the main author. He played a main role in the tasks of compilation and labelling of the corpus, in the development and adaptation of the linguistic features to the domain of misogyny and hate-speech, and the improvement of the tool UMUCorpusClassifier. He wrote the first draft of the article and contributed to its subsequent review. He also made the software to compare the results.

TABLE 3.3: Details of the third publication that compose this PhD Thesis

Key	Value
Title	Psychographic traits identification based on political ideology: An author analysis study on Spanish politicians' tweets posted in 2020
Authors	José Antonio García-Díaz, Ricardo Colomo-Palacios, Rafael Valencia-García
Type	Journal
Journal	Future Generation Computer Systems
Impact Factor	7.187
Publisher	Elsevier
Pages	59–74
Volume	130
Year	2022
Issue	SI: Future-Generation Personality Prediction From Digital Footprints
Month	May
ISSN	0167-739X
DOI	https://doi.org/10.1016/j.future.2021.12.011
URL	https://www.sciencedirect.com/science/article/pii/S0167739X21004921
State	Published

3.3 Psychographic traits identification based on political ideology: An author analysis study on Spanish politicians' tweets posted in 2020

3.3.1 Abstract

In general, people are usually more reluctant to follow advice and directions from politicians who do not have their ideology. In extreme cases, people can be heavily biased in favour of a political party at the same time that they are in sharp disagreement with others, which may lead to irrational decision making and can put people's lives at risk by ignoring certain recommendations from the authorities. Therefore, considering political ideology as a psychographic trait can improve political micro-targeting by helping public authorities and local governments to adopt better communication policies during crises. In this work, we explore the reliability of determining psychographic traits concerning political ideology. Our contribution is twofold. On the one hand, we release the PoliCorpus-2020, a dataset composed by Spanish politicians' tweets posted in 2020. On the other hand, we conduct two authorship analysis tasks with the aforementioned dataset: an author profiling task to extract demographic and psychographic traits, and an authorship attribution task to determine the author of an anonymous text in the political domain. Both experiments are evaluated with several neural network architectures

grounded on explainable linguistic features, statistical features, and state-of-the-art Transformers. In addition, we test whether the neural network models can be transferred to detect the political ideology of citizens. Our results indicate that the linguistic features are good indicators for identifying fine-grained political affiliation, they boost the performance of neural network models when combined with embedding-based features, and they preserve relevant information when the models are tested with ordinary citizens. Besides, we found that lexical and morphosyntactic features are more effective on author profiling, whereas stylometric features are more effective in authorship attribution.

3.3.2 Author's contribution

The PhD student, José Antonio García-Díaz, is the main author. He compiled the dataset and did the development of the software and the experiments, with special emphasis on improving a system of models from ensembles. In this article, the PhD student played a main role regarding dataset decisions, techniques used, and the evaluation of these. Besides, he also wrote the article and its subsequent revision.

3.4 Compilation and evaluation of the Spanish SatiCorpus 2021 for satire identification using linguistic features and transformers

TABLE 3.4: Details of the forth publication that compose this PhD Thesis

Key	Value
Title	Compilation and evaluation of the Spanish SatiCorpus 2021 for satire identification using linguistic features and transformers
Authors	José Antonio García-Díaz, Rafael Valencia-García
Type	Journal
Journal	Complex & Intelligent Systems
Impact Factor	4.927
Publisher	Springer International Publishing
Pages	1–14
Year	2022
Month	January
ISSN	2199-4536
DOI	https://doi.org/10.1007/s40747-021-00625-1
URL	https://link.springer.com/article/10.1007/s40747-021-00625-1
State	Published

3.4.1 Abstract

Satirical content on social media is hard to distinguish from real news, misinformation, hoaxes or propaganda when there are no clues as to which medium these news were originally written in. It is important, therefore, to provide Information Retrieval systems with mechanisms to identify which results are legitimate and which ones are misleading. Our contribution for satire identification is twofold. On the one hand, we release the Spanish SatiCorpus 2021, a balanced dataset that contains satirical and non-satirical documents. On the other hand, we conduct an extensive evaluation of this dataset with linguistic features and embedding-based features. All feature sets are evaluated separately and combined using different strategies. Our best result is achieved with a combination of the linguistic features and BERT with an accuracy of 97.405%. Besides, we compare our proposal with existing datasets in Spanish regarding satire and irony.

3.4.2 Author's contribution

The PhD student, José Antonio García-Díaz, is the main author. He developed all the tasks related to the compilation of the datasets and the author profiling and authorship attribution tasks. He also developed the feature extraction systems for contextual and non-contextual embeddings, and an automatic system to perform a hyper-parameter optimisation process with deep neural networks.

Chapter 4

Conclusions and promising research lines

4.1 Lessons Learned and Conclusions

In this doctoral thesis, we have shown the development and evaluation of a set of linguistic features for Spanish that have proven to be effective in automatic classification tasks. These features are extracted with UMUTextStats, a tool that has been developed during this doctoral thesis and that is available for the research community.

Specifically, two research hypotheses were raised during this thesis. First, if the inclusion of the linguistic features improves the performance of automatic text classification systems in Spanish, and second, if the inclusion of linguistic features can provide interpretability to the models.

For the first research hypothesis we have shown that the linguistic features can be combined easily with state-of-the-art Transformers or traditional machine-learning models, outperforming the results achieved separately. It is worth mentioning that the performance of the linguistic features depends considerably on the task and the domain applied. For example, the results achieved in author analysis task were more promising than in other classification tasks. Moreover, as some of the feature sets relies on stylometric features that are language independent, and that a large portion of the morphosyntactic features is extracted with Stanza [78], we have applied successfully these features in other languages such as Tamil, Hindi or Marathi, among others. This improvement of the performance along with the usage of the linguistic features in different languages, have allowed us to participate in several shared tasks proposed in international workshops such as IberLEF or SemEval, among others. In the majority of the tasks in which we have participated, we achieve competitive results. For example, in Spanish, we ranked 6th in EmoEvalEs 2021 task, or 1st position in two subtasks in MeOffendEs 2021. Besides, we achieved the better score in the Funniness Score Prediction task in HaHa 2021.

For the second research hypothesis, we have obtained the correlation with the Mutual Information measure of the linguistic features with the target class in several domains, including infodemiology, hate-speech and misogyny detection, or emotion analysis, to name but a few. For instance, we found a strong correlation between lexical and morphosyntactic features in author profiling, whereas these kinds of features were less important for conducting authorship attribution. However, stylometric features are more relevant for this particular task. In this sense, we also found that linguistic features related to correction and style are useful for detecting misogyny. Specifically, we found that a number of misspellings were relevant in two Spanish corpora regarding misogyny identification: the Spanish MisoCorpus 2020 and the AMI 2018 dataset. However, this correlation found was smaller in other hate-speech corpora. Similarly, we found that argumentative discourse markers appeared frequently as positive discriminatory linguistic features for misogyny identification. Regarding satire identification, we observed that the features from the linguistic category of correction and style were the most relevant ones. We found, surprisingly, that the number of orthographic errors were more common in non satirical documents.

To summarise the overall work conducted within this doctoral thesis, we report in Table 4.1 a list of all the publications derived from this doctoral thesis, including research articles, workshops, proceedings, and doctoral symposiums. Apart from these publications, and so far this year, we have participated in two SemEval shared tasks regarding sarcasm detection (iSarcasm 2022) [1] and Multimedia Automatic Misogyny Identification [29] (MAMI 2022) and four shared tasks focused on Tamil and English concerning Language Technology for Equality, Diversity, Inclusion (LT-EDI, ACL 2022): (1) DepSign LT-EDI [87], focused on detecting depression signs on social networks; (2) abusive comment detection in Tamil [76]; (3) identification of homophobic and transphobic in comments from YouTube; and (4) emotion detection in Tamil [16].

Year	Type	Research
2018	Workshop	García-Díaz, J. A., Salas-Zárate, M. P., Hernández-Alcaraz, M. L., Valencia-García, R., & Gómez-Berbís, J. M. (2018, March). Machine learning based sentiment analysis on spanish financial tweets. In World Conference on Information Systems and Technologies (pp. 305-311). Springer, Cham. [44]

2018	Workshop	García-Díaz, José Antonio, Oscar Apolinario-Arzube, José Medina-Moreira, José Omar Salavarria-Melo, Katty Lagos-Ortiz, Harry Luna-Aveiga, and Rafael Valencia-García. "Opinion mining for measuring the social perception of infectious diseases. an infodemiology approach." In International Conference on Technologies and Innovation, pp. 229-239. Springer, Cham, 2018. [45]
2018	Workshop	García-Díaz, J. A., Apolinario-Arzube, Ó., Medina-Moreira, J., Luna-Aveiga, H., Lagos-Ortiz, K., & Valencia-García, R. (2018, November). Sentiment Analysis on Tweets related to infectious diseases in South America. In Proceedings of the Euro American Conference on Telematics and Information Systems (pp. 1-5). [46]
2019	Doctoral Symposium	Extracting Spanish Linguistic Features for Natural Language Processing tasks. Proceedings of the Doctoral Symposium of the XXXV International Conference of the Spanish Society for Natural Language Processing (2019). (pp. 32-37)
2020	Workshop	García-Díaz, J. A., Almela, Á., & Valencia-García, R. (2020). UMUTeam at TASS 2020: Combining Linguistic Features and Machine-learning Models for Sentiment Classification. In IberLEF@ SEPLN (pp. 187-196). [31]
2020	Workhop	García-Díaz, J. A., & Valencia-García, R. (2020). UMUTeam at MEX-A3T'2020: Detecting Aggressiveness with Linguistic Features and Word Embeddings. In IberLEF@ SEPLN (pp. 287-292). [40]
2020	Workshop	García-Díaz, J. A., & Valencia-García, R. (2020). UMUTeam at AI-SOCO'2020: Source Code Authorship Identification based on Character N-Grams and Author's Traits. In FIRE (Working Notes) (pp. 717-726). [38]
2020	Doctoral Symposium	Using Linguistic Features for Improving Automatic Text Classification Tasks in Spanish
2020	Workshop	García-Díaz, J. A., Almela, Á., Alcaraz-Mármol, G., & Valencia-García, R. (2020). UMUCorpusClassifier: Compilation and evaluation of linguistic corpus for Natural Language Processing tasks. Procesamiento del Lenguaje Natural, 65, 139-142. [47]

2020	Workshop	García-Díaz, J. A., Apolinario-Arzube, O., & Valencia-García, R. (2020, October). Evaluating Pre-trained Word Embeddings and Neural Network Architectures for Sentiment Analysis in Spanish Financial Tweets. In Mexican International Conference on Artificial Intelligence (pp. 167-178). Springer, Cham. [32]
2020	Article	Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in Latin America [33]
2021	Article	Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings [42]
2021	Workshop	García-Díaz, J. A., Colomo-Palacios, R., & Valencia-García, R. (2021). UMUTeam at EXIST 2021: Sexist Language Identification based on Linguistic Features and Transformers in Spanish and English. [36]
2021	Workshop	Garcıa-Dıaz, J. A., & Valencia-Garcıa, R. (2021). UMUTeam at HAHA 2021: Linguistic Features and Transformers for Analysing Spanish Humor. The What, the How, and to Whom. In Proceedings of the Iberian Languages Evaluation Forum (Iber-LEF 2021), CEUR Workshop Proceedings, Málaga, Spain (Vol. 9). [39]
2021	Workshop	García-Díaz, J. A., Colomo-Palacios, R., & Valencia-Garcia, R. (2021). UMUTeam at EmoEvalEs 2021: Emosjon Analysis for Spanish based on Explainable Linguistic Features and Transformers. [35]
2021	Workshop	García-Dıaz, J. A., Jiménez-Zafra, S. M., & Valencia-Garcıa, R. (2021). Umuteam at meoffendes 2021: Ensemble learning for offensive language identification using linguistic features, fine-grained negation and transformers. In Proceedings of the Iberian Languages Evaluation Forum (Iber-LEF 2021), CEUR Workshop Proceedings. [30]
2021	Doctoral Symposyum	Evaluation of Linguistic Features Separately or Combined with Transformers for Solving Automatic Text Classification Tasks in Spanish

2021	Workshop	García-Díaz, J. A., & Valencia-García, R. (2021, August). UMUTeam at SemEval-2021 Task 7: Detecting and Rating Humor and Offense with Linguistic Features and Word Embeddings. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021) (pp. 1096-1101). [41]
2022	Article	García-Díaz, J. A., & Valencia-García, R. (2022). Compilation and evaluation of the Spanish SatiCorpus 2021 for satire identification using linguistic features and transformers. Complex & Intelligent Systems, 1-14. [37]
2022	Article	García-Díaz, J. A., Jiménez-Zafra, S. M., García-Cumbreras, M. A., & Valencia-García, R. (2022). Evaluating feature combination strategies for hate-speech detection in Spanish using linguistic features and transformers. Complex & Intelligent Systems, 1-22. [43]
2022	Article	García-Díaz, J. A., Colomo-Palacios, R., & Valencia-García, R. (2022). Psychographic traits identification based on political ideology: An author analysis study on Spanish politicians' tweets posted in 2020. Future Generation Computer Systems, 130, 59-74. [34]

TABLE 4.1: Publications derived from this doctoral thesis

4.2 Promising research lines

We will continue with the development and validation of UMUTextStats for different languages and domains. We are currently adapting the taxonomy described here for English and other languages. We expect that the release of the tool to the scientific community make it easier to validate and extend this tool. Besides, we are planning to facilitate the integration of this tool with other NLP tools apart from Stanza. We expect to make it easier to combine and use other NER and PoS models that extend the number of available labels.

As we have observed during the evaluation of the tool, the majority of evaluated texts were short texts from social networks posts. For larger texts or author analysis approaches, we are planning to release the results individually per document and report other metrics, such as the standard deviation.

To use this tool, there are several technical ways. This tool can work using console commands, an API REST, or a graphical interface based on Web technologies. We

have designed an authentication mechanisms in which users can upload different files and obtain the statistics. Besides, this tool can connect to different sources, including plain text files, zipped files, direct input or directly to UMUCorpusClassifier. Our idea is to continue adding more capabilities to the tool and make it easier its usage to people with limited knowledge to command line interfaces.

Next, we focus on the neural network models generated with this tool. Concerning the interpretability of the results, most of the experiments have been conducted from a modal agnostic perspective; that is, evaluating the features outside the machine learning models and using metrics such as Information Gain. However, we need to measure and evaluate the performance of the linguistic features within a neural network. This is specially relevant when the features are combined with embedding based features. Concerning this, we will explore the usage of tools SHAP and LIME [86].

One important drawback of the linguistic features is that the majority of them are contextless, as it happens with other features such as Bag of Words or TF–IDF. To solve this issue, we will evaluate the generation of a small set of linguistic features but extracted token by token rather than from the whole document. This will allow us to use the linguistic features as a sequence and combine them with recurrent neural networks and with attention mechanisms. Moreover, we consider that the linguistic features can be useful in other ways. For example, they can be used for selecting better training, validation, and testing datasets. Instead of performing random sample, the linguistic features can produce better sampling strategies, based on the number of words, length, pronouns, or words that belong to certain category, just to name but a few.

Another promising research line is the improvement of the strategies for combining the linguistic features with other features sets. During the evaluation, we have tested the capabilities of ensemble learning, in which the best characteristics of each model are combined to build more robust models. The strategies evaluated combined the predictions of the individual models; however, modern strategies, such as mixture of experts [23], that rely on the divide-and-conquer principle in which the problem space covers different input regions with different learners, will be evaluated.

We observed that some of the limited results we achieve in some shared tasks are related to class imbalance. One strategy to solve this drawback is data augmentation, which consists in incorporating new samples to the training stage. However, this step is complex. Manual strategies are based on the acquisition of samples from other languages and translate them to the target language. Other strategies consists into creating automatically these new samples using heuristics such as text summarisation. However, these approaches are not always feasible. For example, incorporating translated documents is not particularly useful when

the domain is culturally dependant, as happens with humor and figurative language. We propose the usage of the linguistic features to compare the original training samples with the new ones in data augmentation. Therefore, we can test if the new artificial samples will result in an increment or loss in the performance of the resulting models.

Another utility of the linguistic features is the generation of training-validation-and testing samples of a linguistic corpus. Typical approaches for generating the splits consist into random or stratified sampling. However, the usage of the linguistic features could provide better splits, because we can ensure that all splits contains the same proportion of long or short documents, or a proportional number of words from the same categories.

In case of the UMUCorpusClassifier tool, we will improve it to allow to conduct multi-label classification more easily. Besides, we are incorporating means of obtaining contextual features from the conversation. We will evaluate features related to the time of publication and the popularity of the user in order to observe if they improve or bias the results of hate-speech detectors or author analysis.

We are also extending this tool to extract data from other sources. We are currently implementing a web crawler and we are extracting news from different Spanish newspaper to train Language Models focused on specific domains. In this sense, we will use the UMUTextStat tool to provide useful metrics from these new resources.

Appendix A

Linguistic taxonomy

This appendix contains all the linguistic features organised by categories from UMUTextStats. Categories and written in italics. We include the description of each linguistic category. In order to represent the taxonomy and the relationship between each linguistic category, we include a numeric index.

Index	Feature	Description
1	*phonetics*	
1.1	phonetics-expressive-lengthening	Drawing out or emphasizing a verbalized word, giving it character
2	*morphosyntax*	
2.1	morphosyntax-gender	
2.1.1	morphosyntax-gender-words-feminine	Percentage of grammatically feminine words
2.1.2	morphosyntax-gender-words-masculine	Percentage of grammatically masculine words. In Spanish, masculine gender is the unmarked or inclusive form
2.1.3	morphosyntax-gender-words-neutral	Percentage of grammatically neutral words. Common in articles, abstract concepts, or certain demonstratives among others
2.1.4	morphosyntax-gender-words-common	Grammatically neutral words. Added for languages that do not distinguish between masculine or feminine most of the time but they do distinguish between neutral or non-neutral forms.
2.2	morphosyntax-number	
2.2.1	morphosyntax-number-singular	Counts how many singular words there are in the text
2.2.2	morphosyntax-number-plural	Counts how many plural words there are in the text

2.3	morphosyntax-affixes	
2.3.1	morphosyntax-affixes-suffixes	Counts how many words includes suffixes
2.3.1.1	morphosyntax-affixes-suffixes-nominals	Counts how many words includes nominal suffixes
2.3.1.2	morphosyntax-affixes-suffixes-adjectivizers	Counts how many words includes adjectivizers suffixes
2.3.1.3	morphosyntax-affixes-suffixes-verbalizers	Counts how many words includes verbalizers suffixes
2.3.1.4	morphosyntax-affixes-suffixes-adverbializers	Counts how many words includes adverbializers suffixes
2.3.1.5	morphosyntax-affixes-suffixes-augmentative	Counts how many words includes augmentative suffixes
2.3.1.6	morphosyntax-affixes-suffixes-diminutives	Counts how many words includes diminutive suffixes
2.3.1.7	morphosyntax-affixes-suffixes-despective	Counts how many words includes despective suffixes
2.3.2	morphosyntax-affixes-prefixes	Counts how many words includes prefixes
2.4	morphosyntax-morphology	
2.4.1	morphosyntax-morphology-nouns	Percentage of PoS nouns
2.4.1.1	morphosyntax-morphology-nouns-common	Percentage of PoS common nouns
2.4.1.2	morphosyntax-morphology-nouns-proper	Percentage of PoS proper nouns
2.4.1.3	morphosyntax-morphology-nouns-male	Percentage of male names
2.4.1.4	morphosyntax-morphology-nouns-female	Percentage of female names
2.4.1.5	morphosyntax-morphology-nouns-dates	Sum of the dates percentage
2.4.1.5.1	morphosyntax-morphology-nouns-dates-ymd	Percentage of dates (ymd)
2.4.1.5.2	morphosyntax-morphology-nouns-dates-dmt	Percentage of dates (dmt)
2.4.1.5.3	morphosyntax-morphology-nouns-dates-mdy	Percentage of dates (mdy)
2.4.1.5.4	morphosyntax-morphology-nouns-dates-textual-short	Percentage of dates (short-format)
2.4.1.5.5	morphosyntax-morphology-nouns-dates-textual-long	Percentage of dates (long-format)

2.4.1.5.6	morphosyntax-morphology-nouns-dates-textual-days	Percentage of day names
2.4.1.6	morphosyntax-morphology-nouns-numerals	Percentage of PoS numerals
2.4.2	morphosyntax-morphology-nouns-topics	
2.4.2.1	morphosyntax-morphology-nouns-topics-capitals	Number of capitals
2.4.2.2	morphosyntax-morphology-nouns-topics-countries	Number of countries
2.4.2.3	morphosyntax-morphology-nouns-topics-demonyms	Number of demonyms; that is, to refer to inhabitants of a particular place
2.4.2.4	morphosyntax-morphology-nouns-topics-languages	Number of words that refer to languages
2.4.2.5	morphosyntax-morphology-nouns-topics-honorifics	Number of words that refer to honorifics; that is, titles that convey esteem, courtesy, or respect for position or rank when used in addressing or referring to a person. Sometimes, the term "honorific" is used in a more specific sense to refer to an honorary academic title.
2.4.2.6	morphosyntax-morphology-nouns-topics-colors	Number of words that refer to colors.
2.4.3	morphosyntax-morphology-adjectives	
2.4.3.1	morphosyntax-morphology-adjectives-qualifying	Percentage of qualifying adjectives
2.4.3.2	morphosyntax-morphology-adjectives-ordinals	Percentage of ordinal adjectives
2.4.3.3	morphosyntax-morphology-adjectives-superlative	Percentage of superlative adjectives
2.4.3.4	morphosyntax-morphology-adjectives-despective	Percentage of despective adjectives
2.4.4	morphosyntax-morphology-adverbs	Percentage of adverbs
2.4.4.1	morphosyntax-morphology-adverbs-time	Percentage of adverbs of time
2.4.4.2	morphosyntax-morphology-adverbs-mode	Percentage of adverbs of mode
2.4.4.3	morphosyntax-morphology-adverbs-place	Percentage of adverbs of place

2.4.4.4	morphosyntax-morphology-adverbs-interrogative	Percentage of interrogative adverbs
2.4.4.5	morphosyntax-morphology-adverbs-cause	Percentage of causal adverbs
2.4.4.6	morphosyntax-morphology-adverbs-doubt-or-desire	Percentage of adverbs that expresses doubt or desire
2.4.4.7	morphosyntax-morphology-adverbs-negation	Percentage of negation adverbs
2.4.4.8	morphosyntax-morphology-adverbs-affirmation	Percentage of affirmation adverbs
2.4.4.9	morphosyntax-morphology-adverbs-quantity	Percentage of adverbs that expresses quantity
2.4.4.10	morphosyntax-morphology-adverbs-others	Percentage of adverbs that do not belong to the rest of the categories
2.4.5	morphosyntax-morphology-determiners	
2.4.5.1	morphosyntax-morphology-determiners-articles	Percentage of PoS determiners articles
2.4.5.2	morphosyntax-morphology-determiners-demostrative	Percentage of PoS determiners demostrative
2.4.5.3	morphosyntax-morphology-determiners-emphatic	Percentage of PoS emphatic determiners
2.4.5.4	morphosyntax-morphology-determiners-possessive	Percentage of PoS possessive determiners
2.4.5.5	morphosyntax-morphology-determiners-indefinite	Percentage of PoS indefinite determiners
2.4.5.6	morphosyntax-morphology-determiners-exclamatory	Percentage of PoS exclamatory determiners
2.4.5.7	morphosyntax-morphology-determiners-interrogative	Percentage of PoS interrogative determiners
2.4.5.8	morphosyntax-morphology-determiners-negative	Percentage of PoS negative determiners
2.4.5.9	morphosyntax-morphology-determiners-reciprocal	Percentage of PoS reciprocal determiners
2.4.5.10	morphosyntax-morphology-determiners-relative	Percentage of PoS relative determiners
2.4.5.11	morphosyntax-morphology-determiners-total	Percentage of PoS total determiners
2.4.6	morphosyntax-morphology-pronouns	Personal and impersonal pronouns
2.4.6.1	morphosyntax-morphology-pronouns-personal	Personal pronouns

2.4.6.1.1	morphosyntax-morphology-pronouns-personal-number	Personal pronouns based on number
2.4.6.1.1.1	morphosyntax-morphology-pronouns-personal-number-singular	Personal pronouns in singular
2.4.6.1.1.2	morphosyntax-morphology-pronouns-personal-number-plural	Personal pronouns in plural
2.4.6.1.2	morphosyntax-morphology-pronouns-personal-person	Personal pronouns based on person
2.4.6.1.2.1	morphosyntax-morphology-pronouns-personal-person-first	Personal pronouns based on first person
2.4.6.1.2.2	morphosyntax-morphology-pronouns-personal-person-second	Personal pronouns based on second person
2.4.6.1.2.3	morphosyntax-morphology-pronouns-personal-person-third	Personal pronouns based on third person
2.4.6.1.3	morphosyntax-morphology-pronouns-personal-gender	Personal pronouns based on gender
2.4.6.1.3.1	morphosyntax-morphology-pronouns-personal-gender-male	Male personal pronouns
2.4.6.1.3.2	morphosyntax-morphology-pronouns-personal-gender-female	Female personal pronouns
2.4.6.1.3.3	morphosyntax-morphology-pronouns-personal-gender-neutral	Gender-neutral personal pronouns
2.4.6.1.4	morphosyntax-morphology-pronouns-personal-enclitics	Personal pronouns with enclitics. These pronouns are those attached to verbs so that it can play a role.
2.4.6.2	morphosyntax-morphology-pronouns-impersonal	Impersonal pronouns
2.4.6.3	morphosyntax-morphology-pronouns-indefinite	Indefinite pronouns
2.4.6.4	morphosyntax-morphology-pronouns-relative	Interrogative relative
2.4.6.5	morphosyntax-morphology-pronouns-interrogative	Interrogative pronouns

2.4.6.6	morphosyntax-morphology-pronouns-reciprocal	Reciprocal pronouns
2.4.6.7	morphosyntax-morphology-pronouns-demonstrative	Demonstrative relative
2.4.6.8	morphosyntax-morphology-pronouns-total	Total pronouns
2.4.6.9	morphosyntax-morphology-pronouns-negative	Negative pronouns
2.4.7	morphosyntax-morphology-prepositions	Personal and impersonal pronouns
2.4.7.1	morphosyntax-morphology-prepositions-individual	Percentage of prepositions
2.4.7.2	morphosyntax-morphology-prepositions-locutions	Percentage of prepositional locutions
2.4.8	morphosyntax-morphology-conjunctions	
2.4.8.1	morphosyntax-morphology-conjunctions-coordinating	Percentage of coordinating conjuctions
2.4.8.2	morphosyntax-morphology-conjunctions-subordinating	Percentage of subordinating conjuctions
2.4.9	morphosyntax-morphology-interjections	Percentage of interjections
2.4.10	morphosyntax-morphology-verbs	
2.4.10.1	morphosyntax-morphology-verbs-periphrasis	Percentage of verbal periphrasis. That is, verbal constructions made of two verb forms.
2.4.10.1.1	morphosyntax-morphology-verbs-periphrasis-temporal-aspect	Percentage of verbal periphrasis based on a temporal aspect
2.4.10.1.1.1	morphosyntax-morphology-verbs-periphrasis-temporal-aspect-temporal	Percentage of verbal periphrasis based on a temporal aspect
2.4.10.1.1.1.1	morphosyntax-morphology-verbs-periphrasis-temporal-aspect-temporal-immediate-posteriority	Percentage of verbal periphrasis based on a temporal aspect: immediate posteriority
2.4.10.1.1.1.2	morphosyntax-morphology-verbs-periphrasis-temporal-aspect-temporal-recent-posteriority	Percentage of verbal periphrasis based on a temporal aspect: recent posteriority

2.4.10.1.1.2	morphosyntax-morphology-verbs-periphrasis-temporal-aspect-phase	Percentage of verbal periphrasis based on a phase aspect
2.4.10.1.1.2.1	morphosyntax-morphology-verbs-periphrasis-temporal-aspect-temporal-habit	Percentage of verbal periphrasis for expressing temporal habits
2.4.10.1.1.2.2	morphosyntax-morphology-verbs-periphrasis-temporal-aspect-temporal-imminent	Percentage of verbal periphrasis for expressing imminent actions
2.4.10.1.1.2.3	morphosyntax-morphology-verbs-periphrasis-temporal-aspect-temporal-initial	Percentage of verbal periphrasis for indicating that one action is going to start
2.4.10.1.1.2.4	morphosyntax-morphology-verbs-periphrasis-temporal-aspect-temporal-current	Percentage of verbal periphrases to indicate that an action is occurring
2.4.10.1.1.2.5	morphosyntax-morphology-verbs-periphrasis-temporal-aspect-temporal-ending	Percentage of verbal periphrasis for indicating that one action is going to end
2.4.10.1.1.2.6	morphosyntax-morphology-verbs-periphrasis-temporal-aspect-temporal-transitional	Percentage of verbal periphrasis for indicating that one action is going to change
2.4.10.1.1.2.7	morphosyntax-morphology-verbs-periphrasis-temporal-aspect-temporal-scalar	Percentage of verbal periphrasis for indicating scalar actions
2.4.10.1.2	morphosyntax-morphology-verbs-periphrasis-modals	Percentage of verbal periphrasis for indicating mode
2.4.10.1.2.1	morphosyntax-morphology-verbs-periphrasis-modals-obligation	Percentage of verbal periphrasis for indicating obligation
2.4.10.1.2.2	morphosyntax-morphology-verbs-periphrasis-modals-ability	Percentage of verbal periphrasis for indicating ability
2.4.10.1.2.3	morphosyntax-morphology-verbs-periphrasis-modals-probability	Percentage of verbal periphrasis for indicating probability
2.4.10.1.2.4	morphosyntax-morphology-verbs-periphrasis-modals-certainty	Percentage of verbal periphrasis for indicating certainty
2.4.10.1.2.5	morphosyntax-morphology-verbs-periphrasis-modals-approach	Percentage of verbal periphrasis for indicating approaches

2.4.10.2	morphosyntax-morphology-verbs-number	Percentage of verbs based on numbers
2.4.10.2.1	morphosyntax-morphology-verbs-number-singular	Percentage of singular verbs
2.4.10.2.1.1	morphosyntax-morphology-verbs-number-singular-first	Percentage of singular verbs in first person
2.4.10.2.1.2	morphosyntax-morphology-verbs-number-singular-second	Percentage of singular verbs in second person
2.4.10.2.1.3	morphosyntax-morphology-verbs-number-singular-third	Percentage of singular verbs in third person
2.4.10.2.2	morphosyntax-morphology-verbs-number-plural	Percentage of plural verbs
2.4.10.2.2.1	morphosyntax-morphology-verbs-number-plural-first	Percentage of plural verbs in first person
2.4.10.2.2.2	morphosyntax-morphology-verbs-number-plural-second	Percentage of plural verbs in second person
2.4.10.2.2.3	morphosyntax-morphology-verbs-number-plural-third	Percentage of plural verbs in third person
2.4.10.3	morphosyntax-morphology-verbs-transitivity	
2.4.10.3.1	morphosyntax-morphology-verbs-transitive	Percentage of transitive verbs
2.4.10.3.2	morphosyntax-morphology-verbs-intransitive	Percentage of intransitive verbs
2.4.10.4	morphosyntax-morphology-verbs-inflection	
2.4.10.4.1	morphosyntax-morphology-verbs-inflection-regular	Percentage of regular verbs
2.4.10.4.2	morphosyntax-morphology-verbs-inflection-irregular	Percentage of irregular verbs
2.4.10.5	morphosyntax-morphology-verbs-function	
2.4.10.5.1	morphosyntax-morphology-verbs-function-main	Percentage of main verbs
2.4.10.5.2	morphosyntax-morphology-verbs-function-copulative	Percentage of copulative verbs
2.4.10.5.3	morphosyntax-morphology-verbs-function-auxiliary	Percentage of auxiliary verbs
2.4.10.6	morphosyntax-morphology-verbs-nonfinite	

2.4.10.6.1	morphosyntax-morphology-verbs-nonfinite-infinitive	Percentage of infinitive verbs
2.4.10.6.2	morphosyntax-morphology-verbs-nonfinite-gerund	Percentage of gerund verbs
2.4.10.6.3	morphosyntax-morphology-verbs-nonfinite-participle	Percentage of participle verbs
2.4.10.7	morphosyntax-morphology-verbs-tense	
2.4.10.7.1	morphosyntax-morphology-verbs-tense-past	Counts how many words are past
2.4.10.7.2	morphosyntax-morphology-verbs-tense-present	Counts how many words are present focus
2.4.10.7.3	morphosyntax-morphology-verbs-tense-future	Counts how many words are future focus
2.4.10.8	morphosyntax-morphology-verbs-mode	
2.4.10.8.1	morphosyntax-morphology-verbs-mode-indicative	Counts how many words are in indicative
2.4.10.8.2	morphosyntax-morphology-verbs-mode-subjunctive	Counts how many words are in subjunctive
2.4.10.8.3	morphosyntax-morphology-verbs-imperative	Percentage of imperative verbs
2.4.10.8.4	morphosyntax-morphology-verbs-conditional	Percentage of conditional verbs
2.4.10.9	morphosyntax-morphology-verbs-indicative-simple	Percentage of verbs in indicative simple
2.4.10.9.1	morphosyntax-morphology-verbs-indicative-simple-present	Percentage of verbs in present indicative simple
2.4.10.9.2	morphosyntax-morphology-verbs-indicative-simple-past	Percentage of verbs in past indicative simple
2.4.10.9.3	morphosyntax-morphology-verbs-indicative-simple-future	Percentage of verbs in future indicative simple
2.4.10.9.4	morphosyntax-morphology-verbs-indicative-simple-conditional	Percentage of verbs in conditional indicative simple
2.4.10.9.5	morphosyntax-morphology-verbs-indicative-simple-imperative	Percentage of verbs in imperative simple
2.4.10.10	morphosyntax-morphology-verbs-subjunctive-simple	Percentage of verbs in subjunctive simple

2.4.10.10.1	morphosyntax-morphology-verbs-subjunctive-simple-present	Percentage of verbs in present subjunctive simple
2.4.10.10.2	morphosyntax-morphology-verbs-subjunctive-simple-past	Percentage of verbs in past subjunctive simple
2.4.10.10.3	morphosyntax-morphology-verbs-subjunctive-simple-future	Percentage of verbs in future subjunctive simple
2.4.10.10.4	morphosyntax-morphology-verbs-subjunctive-simple-past-other	Percentage of verbs in other past forms of subjunctive simple
2.4.10.11	morphosyntax-morphology-verbs-indicative-compound	Percentage of compound verbs in indicative
2.4.10.11.1	morphosyntax-morphology-verbs-indicative-compound-present-perfect	Percentage of compound verbs in indicative in present perfect
2.4.10.11.2	morphosyntax-morphology-verbs-indicative-compound-pluperfect	Percentage of compound verbs in indicative in pluperfect
2.4.10.11.3	morphosyntax-morphology-verbs-indicative-compound-past-perfect-tense	Percentage of compound verbs in indicative in past perfect tense
2.4.10.11.4	morphosyntax-morphology-verbs-indicative-compound-future-perfect	Percentage of compound verbs in indicative in future perfect
2.4.10.11.5	morphosyntax-morphology-verbs-indicative-compound-conditional-perfect	Percentage of compound verbs in indicative in conditional perfect
2.4.10.11.6	morphosyntax-morphology-verbs-indicative-compound-past-perfect	Percentage of compound verbs in indicative in past perfect
2.4.10.12	morphosyntax-morphology-verbs-subjunctive-compound	Percentage of compound verbs in subjunctive
2.4.10.12.1	morphosyntax-morphology-verbs-subjunctive-compound-pluperfect	Percentage of compound verbs in pluperfect subjunctive
2.4.10.12.2	morphosyntax-morphology-verbs-subjunctive-compound-future-perfect	Percentage of compound verbs in future perfect in subjunctive

2.4.10.12.3	morphosyntax-morphology-verbs-subjunctive-compound-pluperfect-2	Percentage of compound verbs in future perfect in subjunctive (alternative version)
2.4.10.12.4	morphosyntax-morphology-verbs-subjunctive-compound-past-perfect	Percentage of compound verbs in past perfect in subjunctive
3	*errors*	
3.1	errors-orthographics	Orthographic errors
3.1.1	errors-orthographics-sentences-starting-in-lowercase	Percentage of sentences that start in lowercase
3.1.2	errors-orthographics-misspelled-words	Percentage of misspelled words
3.1.3	errors-orthographics-misspelled-accents-words	Percentage of poorly accented words
3.2	errors-stylistics	Stylistic errors
3.2.1	errors-sentences-starting-with-numbers	Sentences that start with numbers
3.2.2	errors-sentences-starting-with-the-same-word	Sentences that starts with the same word
3.3	errors-performance	Performance errors
3.3.1	errors-performance-duplicated-words	Number of duplicated words
3.3.2	errors-performance-dot-after-exclamation-or-interrogation	Counts how many sentences have a dot after "!" or "?"
3.3.3	errors-performance-two-or-more-consecutive-commas	Two or more consecutive commas
3.3.4	errors-performance-two-or-more-consecutive-periods	Two or more consecutive periods
3.3.5	errors-performance-common-errors	List of common errors in Spanish
3.3.6	errors-performance-redundant-expressions	Percentage of redundant expressions
4	*semantics*	
4.1	semantics-onomatopoeia	List of onomatopoeia; that is, the formation of a word from a sound associated with what is named.
4.2	semantics-euphemisms	List of euphemisms; that is, mild expressions that replaces another that is considered too harsh.

4.3	semantics-dysphemisms	List of dysphemisms; that is, expressions with derogatory connotations.
4.4	semantics-synecdoche	List of synecdoche; that are figures of speech in which a part is made to represent the whole
5	*pragmatics*	
5.1	pragmatics-figurative-language	
5.1.1	pragmatics-figurative-language-hyperboles	Number of expressions that contains hyperboles
5.1.2	pragmatics-figurative-language-idiomatics-expressions	Number of idiomatic expressions
5.1.3	pragmatics-figurative-language-rhetorical-questions	Number of rhetorical questions
5.1.4	pragmatics-figurative-language-verbal-irony	Number of expressions that contains verbal irony
5.1.5	pragmatics-figurative-language-understatements	Number of expressions of understatements
5.1.6	pragmatics-figurative-language-metaphors	
5.1.7	pragmatics-figurative-language-similes	Number of similes
5.2	pragmatics-discourse-markers	
5.2.1	pragmatics-discourse-markers-structuring	Number of discourse markers using for structuring the text
5.2.1.1	pragmatics-discourse-markers-structuring-commenters	Number of discourse markers using for structuring with commenters
5.2.1.2	pragmatics-discourse-markers-structuring-order	Number of discourse markers using for ordering
5.2.2	pragmatics-discourse-markers-connectors	Number of discourse markers for connecting ideas
5.2.2.1	pragmatics-discourse-markers-connectors-additive	Number of discourse markers for adding ideas
5.2.2.2	pragmatics-discourse-markers-connectors-consecutive	Number of discourse markers for connecting ideas

5.2.2.3	pragmatics-discourse-markers-connectors-counter-augmentative	Number of discourse markers for augmentation
5.2.3	pragmatics-discourse-markers-reformers	Number of discourse markers used for reforming
5.2.3.1	pragmatics-discourse-markers-reformers-explanatory	Number of discourse markers used for explaining
5.2.3.2	pragmatics-discourse-markers-reformers-corrective	Number of discourse markers used for correcting
5.2.3.3	pragmatics-discourse-markers-reformers-distance	Number of discourse markers used to move away from the treated fact
5.2.3.4	pragmatics-discourse-markers-reformers-recapitulative	Number of discourse markers used to recapitulative
5.2.4	pragmatics-discourse-markers-argumentative	Number of discourse markers used to argument
5.2.4.1	pragmatics-discourse-markers-argumentative-reinforcement	Number of discourse markers used to reinforce an argument
5.2.4.2	pragmatics-discourse-markers-argumentative-concretion	Number of discourse markers used to specify an argument
5.2.5	pragmatics-discourse-markers-conversational-bookmarks	Number of discourse markers used as bookmarks or pause a conversation
5.3	pragmatics-courtesy-forms	Percentage of courtesy forms
5.3.1	pragmatics-courtesy-forms-greetings	Percentage of expressions used for greetings
5.3.2	pragmatics-courtesy-forms-farewell	Percentage of expressions used for farewell
5.3.3	pragmatics-courtesy-forms-requirements	Percentage of expressions used for requirements
5.3.4	pragmatics-courtesy-forms-general	Percentage of expressions used for requirements
5.3.5	pragmatics-courtesy-forms-condolences	Percentage of expressions used for expressing condolences
6	*stylometry*	
6.1	stylometry-corpus	
6.1.1	stylometry-corpus-length	Counts the length of the text
6.1.2	stylometry-corpus-TTR	Standardized ratios TTR

6.1.3	stylometry-corpus-TTR-standard	TTR (by chunks)
6.1.4	stylometry-corpus-TTR-deviation	TTR (by chunks)
6.1.5	stylometry-corpus-words-count	Counts how many words there are in the text
6.1.6	stylometry-corpus-syllables-count	Counts how many syllables there are in the text
6.1.7	stylometry-corpus-syllables-per-word	Counts how many syllables there per word
6.1.8	stylometry-corpus-words-per-sentence	The average number of words per sentence
6.1.9	stylometry-corpus-uppercase	The average number of words written in uppercase
6.1.10	stylometry-corpus-titlecase	The average number of words written in titlecase
6.1.11	stylometry-corpus-writing-style	
6.1.11.1	stylometry-corpus-readability	Measures the readability within of the text
6.1.11.2	stylometry-corpus-inflesz	Measures the perspicuity within of the text
6.1.12	stylometry-corpus-expressions-quoted	Counts how many quoted expressions
6.1.13	stylometry-corpus-expressions-within-parenthesis	Counts how many expressions within parenthesis
6.1.14	stylometry-corpus-expressions-within-asterisks	Counts how many expressions within asterisks
6.1.15	stylometry-corpus-words-length-avg	The average length of the words
6.1.16	stylometry-corpus-words-longer-6tr	The number of words bigger than 6 characters
6.1.17	stylometry-corpus-words-with-1ltr	The number of words equal to 1 characters
6.1.18	stylometry-corpus-words-with-2ltr	The number of words equal to 2 characters
6.1.19	stylometry-corpus-words-with-3ltr	The number of words equal to 3 characters
6.1.20	stylometry-corpus-words-with-4ltr	The number of word equal to 4 characters

6.1.21	stylometry-corpus-words-with-5ltr	The number of words equal to 5 characters
6.1.22	stylometry-corpus-words-with-6ltr	The number of words equal to 6 characters
6.1.23	stylometry-corpus-words-with-7ltr	The number of words equal to 7 characters
6.1.24	stylometry-corpus-words-with-8ltr	The number of words equal to 8 characters
6.1.25	stylometry-corpus-words-with-9ltr	The number of words equal to 9 characters
6.1.26	stylometry-corpus-words-with-10ltr	The number of words equal to 10 characters
6.1.27	stylometry-corpus-words-with-11ltr	The number of words equal to 11 characters
6.1.28	stylometry-corpus-words-with-12ltr	The number of words equal to 12 characters
6.1.29	stylometry-corpus-words-with-13ltr	The number of words equal to 13 characters
6.1.30	stylometry-corpus-words-with-14ltr	The number of words equal to 14 characters
6.1.31	stylometry-corpus-words-with-15ltr	The number of words equal to 15 characters
6.1.32	stylometry-corpus-words-with-16ltr	The number of words equal to 16 characters
6.1.33	stylometry-corpus-words-with-17ltr	The number of words equal to 17 characters
6.1.34	stylometry-corpus-words-with-18ltr	The number of words equal to 18 characters
6.1.35	stylometry-corpus-words-with-19ltr	The number of words equal to 19 characters
6.1.36	stylometry-corpus-words-with-20ltr	The number of words equal to 20 characters
6.1.37	stylometry-corpus-words-with-21ltr	The number of words equal to 21 characters
6.1.38	stylometry-corpus-words-with-22ltr	The number of words equal to 22 characters
6.1.39	stylometry-corpus-words-with-23ltr	The number of words equal to 23 characters
6.1.40	stylometry-corpus-words-with-24ltr	The number of words equal to 24 characters
6.1.41	stylometry-corpus-words-with-25ltr	The number of words equal to 25 characters

6.2	stylometry-sentences	
6.2.1	stylometry-sentences-count	Counts how many sentences there are in the text
6.2.2	stylometry-sentences-exclamative-percentage	Counts how many exclamative sentences there are in the text
6.2.3	stylometry-sentences-exclamative-percentage-emphasis	Counts how many exclamative sentences there are in the text
6.2.4	stylometry-sentences-interrogative-percentage	Counts how many exclamative sentences there are in the text with more than "?"
6.2.5	stylometry-sentences-interrogative-percentage-emphasis	Counts how many interrogative sentences there are in the text with more than one "?"
6.2.6	stylometry-sentences-quotes-percentage	Counts how many quoted sentences
6.2.7	stylometry-sentences-passive-percentage	Counts how many sentences with pasive voice within the text
6.2.8	stylometry-sentences-conversations	Counts how many sentences simulate a conversation
6.3	stylometry-punctuation-symbols	
6.3.1	stylometry-punctuation-symbols-prime	Number of prime symbols
6.3.2	stylometry-punctuation-symbols-currencies	Number of symbols used as currency symbols
6.3.3	stylometry-punctuation-symbols-apostrophe	Apostrophes, can be used to measure units or in Elision of words
6.3.4	stylometry-punctuation-symbols-brackets-open	Number of open curly brackets
6.3.5	stylometry-punctuation-symbols-curly-brackets-close	Number of closed curly brackets
6.3.6	stylometry-punctuation-symbols-quotation-open	Number of quotation open symbols
6.3.7	stylometry-punctuation-symbols-quotation-close	Number of quotation close symbols
6.3.8	stylometry-punctuation-symbols-colons	Number of colons
6.3.9	stylometry-punctuation-symbols-semicolons	Number of semicolons
6.3.10	stylometry-punctuation-symbols-dashes	Number of dashed

6.3.11	stylometry-punctuation-symbols-low-dashes	Number of low dashed
6.3.12	stylometry-punctuation-symbols-fullstop	Number of fullstops
6.3.13	stylometry-punctuation-symbols-commas	Number of commas
6.3.14	stylometry-punctuation-symbols-open-question	Number of open question symbol
6.3.15	stylometry-punctuation-symbols-open-exclamation	Number of open exclamation symbol
6.3.16	stylometry-punctuation-symbols-space	Number of spaces
6.3.17	stylometry-punctuation-symbols-interdot	Number of interdots
6.3.18	stylometry-punctuation-symbols-percentages	Number of percentages
6.3.19	stylometry-punctuation-symbols-ampersand	Number of ampersands
6.3.20	stylometry-punctuation-symbols-at-sign	Number of at sign symbols
6.3.21	stylometry-punctuation-symbols-backslash	Number of backslash symbols
6.3.22	stylometry-punctuation-symbols-pipe	Number of pipe symbols
6.3.23	stylometry-punctuation-symbols-bullet	Number of bullet symbols
6.3.24	stylometry-punctuation-symbols-caret	Number of caret symbols
6.3.25	stylometry-punctuation-symbols-degree	Number of degree symbols
6.3.26	stylometry-punctuation-symbols-ditto-mark	Number of ditto marks, that expresses that the words or figures above it are to be repeated
6.3.27	stylometry-punctuation-symbols-hash	Number of hash symbols
6.3.28	stylometry-punctuation-symbols-numero-sign	Number of numero signs
6.3.29	stylometry-punctuation-symbols-pilcrow	Number of pilcrow symbol, that are typographical characters that mark the star of a paragraph
6.3.30	stylometry-punctuation-symbols-trademarks	Number of copyright or trademark signs

6.3.31	stylometry-punctuation-symbols-mathematical-signs	Number of mathematical signs
6.3.32	stylometry-punctuation-symbols-line-breaks	Number of line breaks
7	*lexical*	
7.1	lexical-locations	Percentage of words of related to locations
7.2	lexical-organisations	Percentage of words of related to organisations
7.3	lexical-persons	Percentage of words of related to persons
7.4	lexical-others	Percentage of words of related to other named entities
7.5	lexical-animals	Percentage of words of related to animals
7.6	lexical-weapons	Percentage of words of related to weapons
7.7	lexical-food	Percentage of words of related to food
7.8	lexical-jobs	Percentage of words of related to jobs
7.9	lexical-crime	Percentage of words of related to crime
7.10	lexical-personal-money	Percentage of words of related to money
7.11	lexical-personal-religion	Percentage of words of related to religion
7.12	lexical-personal-work	Percentage of words of related to work
7.13	lexical-clothes	Percentage of words of related to clothes
7.14	lexical-body	Percentage of words of related to the body
7.15	lexical-body-male-genitalia	Percentage of words of related to male genitalia
7.16	lexical-body-female-genitalia	Percentage of words of related to female genitalia
7.17	lexical-health	Percentage of words of related to health
7.18	lexical-ingesting	Percentage of words of related to ingesting food and drinks
7.19	lexical-sex	Percentage of words of related to sex
7.20	lexical-death	Percentage of words of related to death

7.21	lexical-home	Percentage of words of related to home
7.22	lexical-social	
7.22.1	lexical-social-inclusive	Percentage of words of inclusive language
7.22.2	lexical-social-analytic	Percentage of words and expressions related to analytic thinking
7.22.3	lexical-social-affiliation	Percentage of words and expressions related to affiliation
7.22.4	lexical-social-achievement	Percentage of words and expressions related to achievement
7.22.5	lexical-social-risk	Percentage of words and expressions related to risk
7.22.6	lexical-social-social-family	Percentage of words and expressions related to family
7.22.7	lexical-social-friendship	Percentage of words and expressions related to friendship
7.22.8	lexical-social-social-female	Percentage of words and expressions related to social female groups and individuals
7.22.9	lexical-social-social-male	Percentage of words and expressions related to social male groups and individuals
7.22.10	lexical-social-cognitive	Percentage of words and expressions related to cognitive processes
7.22.10.1	lexical-social-cognitive-insight	Percentage of words and expressions related to insights
7.22.10.2	lexical-social-cognitive-cause	Percentage of words and expressions related to causes
7.22.10.3	lexical-social-cognitive-discrepancies	Percentage of words and expressions related to discrepancies
7.22.10.4	lexical-social-cognitive-tentativeness	Percentage of words and expressions related to tentativeness
7.22.10.5	lexical-social-cognitive-certainty	Percentage of words and expressions related to certainty
7.22.11	lexical-social-perceptual-processes	Percentage of words and expressions related to feelings and perceptions
7.22.11.1	lexical-social-perceptual-explore	Percentage of words and expressions related to explore and adventure
7.22.11.2	lexical-social-perceptual-surprise	Percentage of words and expressions related to surprise

7.22.11.3	lexical-social-perceptual-feel	Percentage of words and expressions related to the sense of feeling
7.22.11.4	lexical-social-perceptual-hear	Percentage of words and expressions related to the sense of hearing
7.22.11.5	lexical-social-perceptual-see	Percentage of words and expressions related to the sense of seeing
7.22.12	lexical-social-relativity	Percentage of words and expressions related to relative processes
7.22.12.1	lexical-social-relativity-movement	Percentage of words and expressions related to movement
7.22.12.2	lexical-social-relativity-space	Percentage of words and expressions related to space
7.22.12.3	lexical-social-relativity-time	Percentage of words and expressions related to time
8	*psycholinguistic-processes*	
8.1	psycholinguistic-processes-positive	Percentage of positive emotions
8.1.1	psycholinguistic-processes-positive-general	Percentage of positive emotions
8.1.2	psycholinguistic-processes-positive-pleasure	Percentage of positive emotions
8.1.3	psycholinguistic-processes-positive-emoticons	Percentage of positive emotions
8.2	psycholinguistic-processes-negative	Percentage of words related with negative psycholinguistic processes
8.2.1	psycholinguistic-processes-negative-general	Percentage of words related with negative feelings
8.2.2	psycholinguistic-processes-negative-anxiety	Percentage of words related with anxiety
8.2.3	psycholinguistic-processes-negative-anger	Percentage of words related with anger
8.2.4	psycholinguistic-processes-negative-sad	Percentage of words related with sadness
8.2.5	psycholinguistic-processes-negative-emoticons	Percentage of negative emotions
9	*register*	
9.1	register-offensive-speech	
9.1.1	register-offensive-speech	Percentage of words related with offences
9.1.2	register-offensive-speech-soft	Percentage of words related with soft offences

9.1.3	register-offensive-speech-strong	Percentage of words related with strong offences
9.2	register-informal-speech	Average of expressions related to informal speech
9.2.1	register-informal-speech-assent	Expressions related to expressions of approval or agreement.
9.2.2	register-informal-speech-nonfluencies	Percentage of non-fluent words and expressions.
9.2.3	register-informal-speech-swear	Percentage of swear words and expressions.
9.2.4	register-informal-speech-colloquialisms	Percentage of colloquialisms.
9.2.5	register-informal-speech-sms	Percentage of informal speech with a style of short messages on the Internet
9.3	register-polite-language-cultisms	Percentage of cultism
9.4	register-polite-language-latinisms	Percentage of latinisms
10	*social-media*	
10.1	social-media-hashtags	Number of hashtags
10.2	social-media-mentions	Number of mentions
10.3	social-media-mentions-in-the-middle	Number of mentions in the middle of the text
10.4	social-media-urls	Number of hyperlinks
10.5	social-media-jargon	Number of hyperlinks
10.6	social-media-emoticons	Number of emoticons
10.7	social-media-reply-female	Number of replies directed to a female name account
10.8	social-media-reply-male	Number of emoticons

TABLE A.1: UMUTextStats linguistic taxonomy

References

[1] Ibrahim Abu Farha et al. "SemEval-2022 Task 6: iSarcasmEval, Intended Sarcasm Detection in English and Arabic". In: *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics, 2022.

[2] Wasim Ahmed, Peter A Bath, and Gianluca Demartini. "Using Twitter as a data source: An overview of ethical, legal, and methodological challenges". In: *The ethics of online research* (2017).

[3] Angela Almela, Rafael Valencia-García, and Pascual Cantos. "Seeing through deception: A computational approach to deceit detection in Spanish written communication". In: *Linguistic Evidence in Security, Law and Intelligence* 1.1 (2013), pp. 3–12.

[4] Ángela Almela et al. "Developing and Analyzing a Spanish Corpus for Forensic Purposes". In: *Linguistic Evidence in Security, Law and Intelligence* 3 (2019).

[5] ME Aragón et al. "Overview of mex-a3t at iberlef 2020: Fake news and aggressiveness analysis in mexican spanish". In: *Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Malaga, Spain*. 2020.

[6] Flor Miriam Plaza-del Arco et al. "Overview of the EmoEvalEs task on emotion detection for Spanish at IberLEF 2021". In: (2021).

[7] ONAN Aytuğ. "Sentiment analysis on Twitter based on ensemble of psychological and linguistic feature sets". In: *Balkan Journal of Electrical and Computer Engineering* 6.2 (2018), pp. 69–77.

[8] Pedro Balage Filho, Thiago Alexandre Salgueiro Pardo, and Sandra Aluísio. "An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis". In: *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*. 2013.

[9] Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. "Is this Tweet satirical? A computational approach for satire detection in Spanish". In: *Procesamiento del Lenguaje Natural* 55 (2015), pp. 135–142.

[10] IM Barrio-Cantalejo et al. "Validation of the INFLESZ scale to evaluate readability of texts aimed at the patient". In: *Anales del sistema sanitario de Navarra*. Vol. 31. 2. 2008, pp. 135–152.

[11] Valerio Basile et al. "Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter". In: *13th International*

Workshop on Semantic Evaluation. Association for Computational Linguistics. 2019, pp. 54–63.

[12] Peter Boot, Hanna Zijlstra, and Rinie Geenen. "The Dutch translation of the linguistic inquiry and word count (LIWC) 2007 dictionary". In: *Dutch Journal of Applied Linguistics* 6.1 (2017), pp. 65–76.

[13] Cristian Cardellino. *Spanish Billion Words Corpus and Embeddings*. 2019. URL: https://crscardellino.github.io/SBWCE/.

[14] José Cañete et al. "Spanish pre-trained bert model and evaluation data". In: *Pml4dc at iclr* 2020 (2020), p. 2020.

[15] José Cañete. *Compilation of Large Spanish Unannotated Corpora*. May 2019. DOI: 10.5281/zenodo.3247731. URL: https://doi.org/10.5281/zenodo.3247731.

[16] Bharathi Raja Chakravarthi et al. "Findings of the Shared Task on Homophobia Transphobia Detection in Social Media Comments". In: *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics, May 2022.

[17] Ngoni Chipere, David Malvern, and Brian Richards. "Using a corpus of children's writing to test a solution to the sample size problem affecting type-token ratios". In: *Corpora and language learners* (2004), pp. 139–147.

[18] Luis Chiruzzo et al. "Overview of HAHA at IberLEF 2021: Detecting, Rating and Analyzing Humor in Spanish". In: *Procesamiento del Lenguaje Natural* 67 (2021), pp. 257–268.

[19] Alexis Conneau et al. "Unsupervised Cross-lingual Representation Learning at Scale". In: *CoRR* abs/1911.02116 (2019). arXiv: 1911.02116. URL: http://arxiv.org/abs/1911.02116.

[20] Lindsay Grey Cowell and Barry Smith. "Infectious disease ontology". In: *Infectious disease informatics*. Springer, 2010, pp. 373–395.

[21] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[22] S. H. H. Ding et al. "Learning Stylometric Representations for Authorship Analysis". In: *IEEE Transactions on Cybernetics* 49.1 (2019), pp. 107–121. DOI: 10.1109/TCYB.2017.2766189.

[23] Nan Du et al. "GLaM: Efficient Scaling of Language Models with Mixture-of-Experts". In: *arXiv preprint arXiv:2112.06905* (2021).

[24] Paul Ekman. "Lie catching and microexpressions". In: *The philosophy of deception* 1.2 (2009), p. 5.

[25] Gunther Eysenbach. "Infodemiology: The epidemiology of (mis) information". In: *The American journal of medicine* 113.9 (2002), pp. 763–765.

[26] Ali Fadel et al. "Overview of the PAN@ FIRE 2020 task on Authorship Identification of SOurce COde (AI-SOCO)". In: *Proceedings of The 12th meeting of the Forum for Information Retrieval Evaluation (FIRE 2020), CEUR Workshop Proceedings, CEUR-WS. org*. 2020.

[27] Elisabetta Fersini, Enza Messina, and Federico Alberto Pozzi. "Expressive signals in social media languages to improve polarity detection". In: *Information Processing & Management* 52.1 (2016), pp. 20–35.

[28] Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. "Overview of the Task on Automatic Misogyny Identification at IberEval 2018." In: *IberEval@ SEPLN* 2150 (2018), pp. 214–228.

[29] Elisabetta Fersini et al. "SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification". In: *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics, 2022.

[30] JA Garcá-Dıaz, Salud Marıa Jiménez-Zafra, and Rafael Valencia-Garcıa. "Umuteam at meoffendes 2021: Ensemble learning for offensive language identification using linguistic features, fine-grained negation and transformers". In: *Proceedings of the Iberian Languages Evaluation Forum (Iber-LEF 2021), CEUR Workshop Proceedings. CEUR-WS. org*. 2021.

[31] José Antonio García-Díaz, Ángela Almela, and Rafael Valencia-García. "UMUTeam at TASS 2020: Combining Linguistic Features and Machine-learning Models for Sentiment Classification". In: *Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Malaga, Spain*. 2020, pp. 187–196.

[32] José Antonio García-Díaz, Oscar Apolinario-Arzube, and Rafael Valencia-García. "Evaluating Pre-trained Word Embeddings and Neural Network Architectures for Sentiment Analysis in Spanish Financial Tweets". In: *Mexican International Conference on Artificial Intelligence*. Springer. 2020, pp. 167–178.

[33] José Antonio García-Díaz, Mar Cánovas-García, and Rafael Valencia-García. "Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in Latin America". In: *Future Generation Computer Systems* 112 (2020), pp. 641–657.

[34] José Antonio García-Díaz, Ricardo Colomo-Palacios, and Rafael Valencia-García. "Psychographic traits identification based on political ideology: An author analysis study on Spanish politicians' tweets posted in 2020". In: *Future Generation Computer Systems* (2021).

[35] José Antonio García-Díaz, Ricardo Colomo-Palacios, and Rafael Valencia-Garcia. "UMUTeam at EmoEvalEs 2021: Emosjon Analysis for Spanish based on Explainable Linguistic Features and Transformers". In: (2021).

[36] José Antonio García-Díaz, Ricardo Colomo-Palacios, and Rafael Valencia-García. "UMUTeam at EXIST 2021: Sexist Language Identification based on Linguistic Features and Transformers in Spanish and English". In: (2021).

[37] José Antonio García-Díaz and Rafael Valencia-García. "Compilation and evaluation of the Spanish SatiCorpus 2021 for satire identification using linguistic features and transformers". In: *Complex & Intelligent Systems* (2022), pp. 1–14.

[38] José Antonio García-Díaz and Rafael Valencia-García. "UMUTeam at AI-SOCO'2020: Source Code Authorship Identification based on Character N-Grams and Author's Traits." In: *FIRE (Working Notes)*. 2020, pp. 717–726.

[39] José Antonio Garcıa-Dıaz and Rafael Valencia-Garcıa. "UMUTeam at HAHA 2021: Linguistic Features and Transformers for Analysing Spanish Humor. The What, the How, and to Whom". In: *Proceedings of the Iberian Languages Evaluation Forum (Iber-LEF 2021), CEUR Workshop Proceedings, Málaga, Spain*. Vol. 9. 2021.

[40] José Antonio García-Díaz and Rafael Valencia-García. "UMUTeam at MEX-A3T'2020: Detecting Aggressiveness with Linguistic Features and Word Embeddings". In: *Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Malaga, Spain*. 2020, pp. 287–292.

[41] José Antonio García-Díaz and Rafael Valencia-García. "UMUTeam at SemEval-2021 Task 7: Detecting and Rating Humor and Offense with Linguistic Features and Word Embeddings". In: *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. 2021, pp. 1096–1101.

[42] José Antonio García-Díaz et al. "Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings". In: *Future Generation Computer Systems* 114 (2021), pp. 506–518.

[43] José Antonio García-Díaz et al. "Evaluating feature combination strategies for hate-speech detection in Spanish using linguistic features and transformers". In: *Complex & Intelligent Systems* (2022), pp. 1–22.

[44] José Antonio García-Díaz et al. "Machine learning based sentiment analysis on spanish financial tweets". In: *World Conference on Information Systems and Technologies*. Springer. 2018, pp. 305–311.

[45] José Antonio García-Díaz et al. "Opinion mining for measuring the social perception of infectious diseases. an infodemiology approach". In: *International Conference on Technologies and Innovation*. Springer. 2018, pp. 229–239.

[46] José Antonio García-Díaz et al. "Sentiment Analysis on Tweets related to infectious diseases in South America". In: *Proceedings of the Euro American Conference on Telematics and Information Systems*. 2018, pp. 1–5.

[47] José Antonio García-Díaz et al. "UMUCorpusClassifier: Compilation and evaluation of linguistic corpus for Natural Language Processing tasks". In: *Procesamiento del Lenguaje Natural* 65 (2020), pp. 139–142.

[48] Manuel García-Vega et al. "Overview of TASS 2020: Introducing Emotion Detection". In: *Proceedings of TASS* (2020).

[49] Joshua Gaston et al. "Authorship attribution vs. adversarial authorship from a liwc and sentiment analysis perspective". In: *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE. 2018, pp. 920–927.

[50] Edouard Grave et al. "Learning Word Vectors for 157 Languages". In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

[51] Asier Gutiérrez-Fandiño et al. *Spanish Language Models*. 2021. arXiv: 2107.07253 [cs.CL].

[52] Nicholas S Holtzman et al. "Linguistic markers of grandiose narcissism: A LIWC analysis of 15 samples". In: *Journal of Language and Social Psychology* 38.5-6 (2019), pp. 773–786.

[53] Chin-Lan Huang et al. "The development of the Chinese linguistic inquiry and word count dictionary." In: *Chinese Journal of Psychology* (2012).

[54] Salud María Jiménez-Zafra et al. "Negation detection for sentiment analysis: A case study in spanish". In: *Natural Language Engineering* 27.2 (2021), pp. 225–248.

[55] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. "Computational methods in authorship attribution". In: *Journal of the American Society for information Science and Technology* 60.1 (2009), pp. 9–26.

[56] Klaus Krippendorff. "Agreement and information in the reliability of coding". In: *Communication Methods and Measures* 5.2 (2011), pp. 93–112.

[57] Artur Kulmizev et al. "The power of character n-grams in native language identification". In: *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. 2017, pp. 382–389.

[58] Surafel M Lakew et al. "Improving zero-shot translation of low-resource languages". In: *arXiv preprint arXiv:1811.01389* (2018).

[59] Quoc V. Le and Tomás Mikolov. "Distributed Representations of Sentences and Documents". In: *CoRR* abs/1405.4053 (2014). arXiv: 1405 . 4053. URL: http://arxiv.org/abs/1405.4053.

[60] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *CoRR* abs/1907.11692 (2019). arXiv: 1907 . 11692. URL: http://arxiv.org/abs/1907.11692.

[61] Estanislao López-López et al. "LIWC-based sentiment analysis in Spanish product reviews". In: *Distributed Computing and Artificial Intelligence, 11th International Conference*. Springer. 2014, pp. 379–386.

[62] Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

[63] José Luis Martí Ferriol. "Selection and validation of a measurement instrument for readability calculations in patient information leaflets for oncological patiens in Spain". In: (2016).

[64] Roger McHaney, Antuela Tako, and Stewart Robinson. "Using LIWC to choose simulation approaches: A feasibility study". In: *Decision Support*

Systems 111 (2018), pp. 1–12. ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2018.04.002. URL: https://www.sciencedirect.com/science/article/pii/S0167923618300691.

[65] JA Meaney et al. "Semeval 2021 task 7: Hahackathon, detecting and rating humor and offense". In: *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. 2021, pp. 105–119.

[66] Tabea Meier et al. ""LIWC auf Deutsch": The development, psychometrics, and introduction of DE-LIWC2015". In: *PsyArXiv* a (2019).

[67] Tomas Mikolov et al. "Advances in Pre-Training Distributed Word Representations". In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

[68] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[69] Sandip Modha et al. "Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages and Conversational Hate Speech". In: *FIRE: 2021 Forum for Information Retrieval Evaluation, Virtual Event, 13th-17th December 2021*. ACM, 2021.

[70] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[71] Juan Carlos Pereira-Kohatsu et al. "Detecting and monitoring hate speech in Twitter". In: *Sensors* 19.21 (2019), p. 4654.

[72] María del Pilar Salas-Zárate et al. "Automatic detection of satire in Twitter: A psycholinguistic-based approach". In: *Knowledge-Based Systems* 128 (2017), pp. 20–33.

[73] María del Pilar Salas-Zárate et al. "Review of English literature on figurative language applied to social networks". In: *Knowledge and Information Systems* 62.6 (2020), pp. 2105–2137.

[74] A. Piolat et al. "La version française du dictionnaire pour le LIWC : modalités de construction et exemples d'utilisation". In: *Psychologie Française* 56.3 (2011), pp. 145–159. ISSN: 0033-2984. DOI: https://doi.org/10.1016/j.psfr.2011.07.002. URL: https://www.sciencedirect.com/science/article/pii/S0033298411000355.

[75] Flor Miriam Plaza-del-Arco et al. "Overview of the MeOffendEs task on offensive text detection at IberLEF 2021". In: *Procesamiento del Lenguaje Natural* 67.0 (2021). ISSN: 1989-7553.

[76] Ruba Priyadharshini et al. "Findings of the Shared Task on Abusive Comment Detection in Tamil". In: *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics, May 2022.

[77] René T. Proyer and Kay Brauer. "Exploring adult Playfulness: Examining the accuracy of personality judgments at zero-acquaintance and an LIWC analysis of textual information". In: *Journal of Research in Personality* 73 (2018), pp. 12–20. ISSN: 0092-6566. DOI: https://doi.org/10.1016/j.jrp.2017.10.002. URL: https://www.sciencedirect.com/science/article/pii/S0092656617301058.

[78] Peng Qi et al. "Stanza: A python natural language processing toolkit for many human languages". In: *arXiv preprint arXiv:2003.07082* (2020).

[79] Nairán Ramírez-Esparza et al. "La psicología del uso de las palabras: Un programa de computadora que analiza textos en español". In: *Revista mexicana de psicología* 24.1 (2007), pp. 85–99.

[80] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2019. URL: https://arxiv.org/abs/1908.10084.

[81] Philip Resnik, Anderson Garron, and Rebecca Resnik. "Using topic modeling to improve prediction of neuroticism and depression in college students". In: *Proceedings of the 2013 conference on empirical methods in natural language processing*. 2013, pp. 1348–1353.

[82] Miguel Ángel Rodríguez-García et al. "Ontology-based annotation and retrieval of services in the cloud". In: *Knowledge-based systems* 56 (2014), pp. 15–25.

[83] Francisco Rodríguez-Sánchez et al. "Overview of exist 2021: sexism identification in social networks". In: *Procesamiento del Lenguaje Natural* 67 (2021), pp. 195–207.

[84] Jeremy Rogers and Steve Gunn. "Identifying feature relevance using a random forest". In: *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection"*. Springer. 2005, pp. 173–184.

[85] Elena Rudkowsky et al. "More than bags of words: Sentiment analysis with word embeddings". In: *Communication Methods and Measures* 12.2-3 (2018), pp. 140–157.

[86] Yves Rychener et al. "Sentence-Based Model Agnostic NLP Interpretability". In: *CoRR* abs/2012.13189 (2020). arXiv: 2012.13189. URL: https://arxiv.org/abs/2012.13189.

[87] Kayalvizhi Sampath et al. "Findings of the Shared Task on Detecting Signs of Depression from Social Media". In: *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics, May 2022.

[88] Kim Schouten and Flavius Frasincar. "Survey on aspect-level sentiment analysis". In: *IEEE Transactions on Knowledge and Data Engineering* 28.3 (2015), pp. 813–830.

[89] Lynn Marie Schriml et al. "Disease Ontology: a backbone for disease semantic integration". In: *Nucleic acids research* 40.D1 (2012), pp. D940–D946.

[90] Michele Settanni and Davide Marengo. "Sharing feelings online: studying emotional well-being via automated text analysis of Facebook posts". In: *Frontiers in psychology* 6 (2015), p. 1045.

[91] Yla R Tausczik and James W Pennebaker. "The psychological meaning of words: LIWC and computerized text analysis methods". In: *Journal of language and social psychology* 29.1 (2010), pp. 24–54.

[92] Shirui Wang, Wenan Zhou, and Chao Jiang. "A survey of word embeddings based on deep learning". In: *Computing* 102.3 (2020), pp. 717–740.

[93] T Wilson, SA Raaijmakers, et al. "Comparing word, character, and phoneme n-grams for subjective utterance recognition". In: *INTERSPEECH 2008-9th Annual Conference of the International Speech Communication Association, 22-26 September 2008, Brisbane, QLD, Canada*. 2008, p. 1614.