Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6035–6044
Marseille, 20-25 June 2022

UMUTextStats: A linguistic feature extraction tool for Spanish

José Antonio García-Díaz, Pedro José Vivancos-Vicente, Ángela Almela, Rafael Valencia-García
Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain
Vócali Sistemas Inteligentes S.L., Parque Científico de Murcia,
Complejo de Espinardo, 30100, Spain
Facultad de Letras, Universidad de Murcia, Campus de La Merced, 30001, Murcia, Spain
{joseantonio.garcia8, angelalm, valencia}@um.es
pedro.vivancos@vocali.net

Abstract

Feature Engineering consists in the application of domain knowledge to select and transform relevant features to build efficient machine learning models. In the Natural Language Processing field, the state of the art concerning automatic document classification tasks relies on word and sentence embeddings built upon deep learning models based on transformers that have outperformed the competition in several tasks. However, the models built from these embeddings are usually difficult to interpret. On the contrary, linguistic features are easy to understand, they result in simpler models, and they usually achieve encouraging results. Moreover, both linguistic features and embeddings can be combined with different strategies which result in more reliable machine-learning models. The de facto tool for extracting linguistic features in Spanish is LIWC. However, this software does not consider specific linguistic phenomena of Spanish such as grammatical gender and lacks certain verb tenses. In order to solve these drawbacks, we have developed UMUTextStats, a linguistic extraction tool designed from scratch for Spanish. Furthermore, this tool has been validated to conduct different experiments in areas such as infodemiology, hate-speech detection, author profiling, authorship verification, humour or irony detection, among others. The results indicate that the combination of linguistic features and embeddings based on transformers are beneficial in automatic document classification.

Keywords: Linguistic Features, Feature Engineering, Natural Language Processing

1. Introduction

Online communication has opened a new way of exploring how people communicate and interact with each other. Recent trends in Natural Language Processing, Artificial Intelligence and Information Retrieval have eased the analysis of large amounts of data that can be used for conducting automatic document classification tasks, such as Sentiment Analysis (SA), with applications in marketing in order to gain faster insights about brands, products and services. Other popular applications of automatic document classification are authorship attribution (AA), which can help uncover anonymous threats, plagiarism detection, and cyberbullying and hate-speech (HS) detection, building safer social environments.

The state of the art concerning automatic document classification is focused on pre-trained embeddings that have been learned with large corpora. These models learn to represent words as vectors of fixed size and arrange them within the latent space, clustering words that have a close relationship. These models are learned based on unsupervised tasks, such as Masked Language Modelling (MLM). Moreover, transformers and attention mechanisms have enabled the learning of contextual embeddings, in which words are aware of surrounding ones, solving problems partially related to polysemy and word disambiguation. Word embeddings can be used as input of other Machine Learning models to perform downstream tasks of automatic document classification. These embeddings can capture rich semantic and lexical variety; however, there are linguistic phenomena that can help to classify documents that can go unnoticed with word embeddings. For example, using uppercase words or vocabulary range may indicate relevant findings that may not be detected easily with embeddings.

Linguistic Features (LF) are features that characterise uses of language in a text. LF can refer to grammatical aspects of a text, analysing how words and sentences are related. Besides, LF can show prosodic features related to stress and intonation or searching for specific lexicons that can indicate different demographic or psychographic features of the authors. Although the performance of the LF is, generally, poorer than state-of-the-art word embeddings, we argue that the LF are easier to interpret and result in simpler models and showing competitive results. Furthermore, LF and word embeddings can be easily combined using different strategies such as Knowledge Integration (KI) or Ensemble Learning (EL), improving the results achieved separately.

We present UMUTextStats, a linguistic feature extraction tool designed for Spanish. This tool can extract a vector made up of the percentages of words and expressions that fit into a series of linguistic categories and features. This tool is based on Linguistic Inquiry and Word Count (LIWC) (Tausczik and Pennebaker, 2010) in an attempt to resolve the deficiencies found during

the translation of LIWC to Spanish (Ramírez-Esparza et al., 2007). In addition, we show an extensive evaluation of these features in different tasks and domains such as infodemiology, hate-speech identification, and authorship attribution, among others.

The rest of the manuscript is organised as follows. First, Section 2 describes LIWC, as it is the base of our proposal, and other linguistic feature extraction tools and their applications. Following this, Section 3 describes the architecture of UMUTextStats and how it classifies the LF into categories and dimensions. The reader can find a detailed list of the experimentation conducted with UMUTextStats in Section 4 and, finally, the conclusions and promising further research lines in Section 5.

2. State of the art

LIWC is the de facto standard for extracting LF (Tausczik and Pennebaker, 2010). It is widely used in text analysis and automatic document classification tasks. Its behaviour is quite straightforward, that is, it reads a collection of texts and counts the percentage of words according to certain psychologically relevant categories that are arranged hierarchically. The way in which LIWC works is based on lexicons. That is to say, each dimension is composed by a dictionary of terms.

The dictionaries of the previous versions of LIWC were built manually and validated by several human annotators. The last known version of LIWC was released in 2015. It contains almost 6,400 words in its master dictionary (1), (2) word stems, and (3) emoticons. It is worth mentioning that one word can belong to multiple categories. Saying more, the current version of LIWC extends these dictionaries automatically using distant methods with large corpora. For this, the authors of LIWC test if each of the words in the dictionaries are related to the rest of the words within the same dictionary in a statistically significant way. This process allowed to evaluate if each word in a particular category is indeed related to the rest of the words in the same dictionary. Besides, this step helps identify other words that passed unnoticed in the manual compilation process.

LIWC distinguishes mainly between two types of words: content words and function words. Content words are nouns, verbs, adjectives, and adverbs. Content words convey the content of the communication. Style or function words are pronouns, prepositions, articles, and conjunctions. Besides, some forms of auxiliary verbs can be considered as style words. Most of the words in a language are content words; however, style words make up over half of the words in a text. Under a psychological point of view, style words reflect how people communicate, whereas content words convey the intention to communicate.

LIWC has a Spanish version validated in (Ramírez-Esparza et al., 2007). It is worth mentioning that, according to the LIWC manual[1], the dictionaries of the Spanish version of LIWC are based on the versions of LIWC 2001 and 2007, but not on the 2015 version. The Spanish version of LIWC has been used in several domains, such as opinion mining (del Pilar Salas-Zárate et al., 2014; López-López et al., 2014; García-Díaz et al., 2018), depression assessment (Ramirez-Esparza et al., 2008), deceit detection (Almela et al., 2013), and satire identification (del Pilar Salas-Zárate et al., 2017). However, during the translation of LIWC, the present authors identified some drawbacks concerning some linguistic issues between the English and the Spanish version. For instance, the fact that some grammatical phenomena of Spanish were not considered, which results in the loss of grammatical gender identification or the lack of many verb conjugations.

3. System architecture

UMUTextStats is a linguistic feature extraction system designed for Spanish. Like LIWC, this system can extract 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 deficiencies found in the Spanish translation of LIWC (Ramírez-Esparza et al., 2007). Apart from these issues, there are other inconveniences that are shared by the Spanish and the English version of LIWC: (1) its arbitrary design of the linguistic categories and features, in which the words that belong to certain categories were selected by a limited number of human annotators; and (2) the fact that LIWC is based principally on simple term-count, so it does not consider the context of a word (that is, the surrounding words that can alter the meaning of a single word).

To solve the aforementioned drawbacks, for the design of the UMUTextStats tool we have created a tree-based structure for defining and arranging the LF within categories and dimensions. Besides, we have included several mechanisms to extend UMUTextStats with new dimensions based on custom dictionaries, regular expressions, and a wide variety of performance errors or specific argot used in social networks. We have also developed a new system for extracting all Spanish verb tenses, including compound verbs and periphrases.

Figure 1 depicts the system architecture of UMUTextStats. Below, we describe each module in detail.

3.1. Source Resolver

This module is responsible for obtaining the source documents to extract the LF. The current version of UMUTextStats accepts the insertion of text directly using the console, plain text, zip files with several single files, and datasets extracted using the UMUCorpusClassifier tool (García-Díaz et al., 2020a).

[1] https://mcrc.journalism.wisc.edu/files/2018/04/Manual_LIWC.pdf

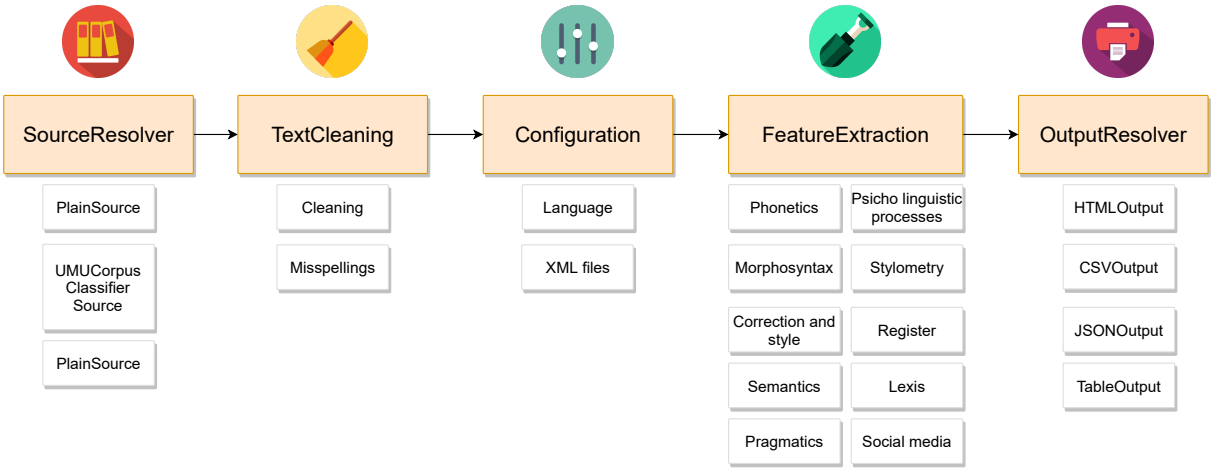

Figure 1: System architecture of UMUTextStats

3.2. Text Cleaning

The source documents are encapsulated in objects of type `TextAsset`. These objects store the text of the document but also caches cleaned and normalised versions of the texts. We find an optional step at this point, which is that the source files can already have a cleaned version. This is useful, for instance, when a custom cleaned version of the text is needed.

If a cleaned version of the text is not provided, UMUTextStats performs a cleaning process that involves: (1) removing blank lines, (2) striping HTML tags, (3) removing URLs, mentions and emojis, (4) removing letter elongations, and (5) converting the text to their lowercase form. Optionally, it is possible to use the PSPELL library to fix misspellings. For this purpose, UMUTextStats analyses each word in isolation and replaces the misspellings with the best suggestion only if the text similarity between the two words is higher than a certain threshold.

3.3. Configuration

The Configuration module is the core of UMUTextStats. It is an XML file that defines the LF and categories in a tree-based structure. It is worth mentioning that the design of UMUTextStats considers maintainability and extensibility as first-citizen requirements. UMUTextStats comes with a series of predefined classes that include the usage of lexicons or regular expressions as typical use cases. Below, we describe the available classes:

- **Dictionary-based features**. This class allows defining new dimensions based on keyword lists. The keywords can be regular expressions. It is possible, however, to configure this dimension disabling regular expressions to speed up the process. Besides, dictionary-based features have other options. For example, it is possible to define counterexamples. The benefit of using counterexamples is that it is easier to define few general regular expressions, and then, to list the exceptions.

- **Verb-based features**. This class is similar to the dictionary-based features class, as verbs are stored in plain text files. However, the large number of verbs make the usage of dictionary-based dimensions impractical. Verb-based features are optimised to identify verbs in O(1). In addition, it is possible to use a custom word separator to consider auxiliary verbs as part of a matching.

- **Sentence per dictionary-based features**. This class calculates the number of sentences that match a regular expression. This class is useful, for example, to obtain the number of sentences that use passive voice of verbs.

- **Features based on enclitic personal pronouns**. This class captures personal pronouns with clitics. In Spanish, the pronouns *la*, *lo*, *le*, *los*, *las*, and *les* are clitic. They indicate direct or indirect third-person pronominal object.

- **Perspicuity-based features**. This class obtains the Degrees of Perspicuity score, according to Flesch-Szigriszt (Barrio-Cantalejo et al., 2008).

- **Readability-based features**. This class obtains the readability score, based on Fernández-Huerta formula (Martí Ferriol, 2016).

- **Grammatical-gender-based features**. This class captures Spanish grammatical gender. It extends the dictionary-based features class but relies on a list of basic rules for obtaining Spanish grammatical gender combined with a list of counterexamples. In addition, this class considers only certain words based on their Part-of-Speech (PoS) category. This way, it is easier to discard rare and made-up words.

- **STTR-based features**. This class allows to calculate the Standardised Type/Token Ratio (STTR), which is the ratio between the total unique words between the total of the words of a text. For long pieces of text, it is possible to calculate this ratio in chunks of N words. It is possible to use the standard deviation of this metric as a linguistic feature.

- **Features based on misspellings**. This class counts the number of misspellings. This class relies on the PSPELL library to capture misspellings, next it checks if the first suggestion of PSPELL is the same word.

- **Features based on wrong accentuation usage**. This class is like the Error-misspelling-based features class, but it is focused on counting misspellings based on wrong usage of the written accents. For this purpose, this class also relies on the PSPELL library to capture misspellings, next it checks if the first suggestion of PSPELL is the same word.

- **PoS-Tagging-based features**. This class counts the number of words that match a specific PoS category. PoS categories are calculated using the Stanza Library (Qi et al., 2020).

- **NER-based features**. This class counts the number of words that match certain NER (Named Entity Recognition) category. Like the PoS-Tagging-based features class, UMUTextStats relies on Stanza (Qi et al., 2020). However, its Spanish model only considers four categories: Person, Location, Organisation, and Miscellaneous.

- **Features based on two or more equal words**. This class detects two or more similar words together. Although this does not have to be an error, it can be an indication of a lack of attention when reviewing a text.

- **Features based on incorrect capitalisation**. This class accounts for the number of sentences that starts in lowercase.

- **Pattern-based features**. This class includes all the occurrences matching certain regular expressions. For example, using the following regular expression `(.)1{3,}<`, it is possible to capture expressive lengthening, that is, emphasising a verbalised word.

- **Typography-based features**. This class allows detecting 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 loud volume of the voice.

- **Composite-based features**. This class allows obtaining certain LF using the Composite Pattern. The aim of this pattern is to define a new dimension based on averaging, adding, subtracting, calculating the maximum or the minimum.

- **Features based on word length**. This class includes the number of words that match or exceed a specific threshold. For this purpose, it is possible to configure word length and equality.

- **Features based on word average length**. This class calculates the average length of all the words in the input.

- **Features based on words per sentence**. This class calculates the number of words per sentence.

- **Features based on unique words**. This class calculates the number of unique words.

- **Features based on syllables per word**. This class calculates the number of syllables per word.

- **Features based on character count**. This class counts the number of a list of specific characters. This class is useful, for example, to capture characters that can be represented with different symbols, such as quotes, currencies, prime symbols, or brackets, among others.

- **Features based on sentences starting with the same word**. This class counts the number of sentences that starts with the same word. This is a custom class to capture certain stylistic errors.

- **Features based on sentences starting with numbers**. This class counts the number of sentences that start with a number, which is considered as a poor writing style.

- **Features based on Twitter's Reply to**. This class is specific for the Social Network category. It determines if a certain text (usually, a Tweet from Twitter) is a response to a specific user based on a list of names.

For example, in Listing 1, we show how to create a LF that captures how many exclamatory sentences are in a document. As can be observed, this dimension relies on the `uses the Pattern-based features` class, so it is easy to define it using a regular expression that matches sentences that end with two or more exclamation marks.

Listing 1: An example of a feature in the configuration

```
<dimension>
    <key>stylometry-sentences-
        exclamatory-percentage-
        emphasis</key>
    <class>PatternDimension</class>
```

```
    <description>Counts how many
        exclamatory sentences there
        are in the text</description>
    <pattern>[\p{L}\p{N}\s]+[\!]{2,}
        </pattern>
    <separator>by-sentence</
        separator>
</dimension>
```

Besides, some classes have common options. For example, dictionary and pattern-based features allow for defining the separator. Usually, the texts are divided into words, but it is also possible to separate texts into sentences or to apply the regular expression to the whole text. Sentence splitting is useful, for example, to count exclamatory sentences. Another option is to indicate if the dimension must obtain the raw count or the percentage. It is worth noting that the percentage is related to the separator employed. For example, in the aforementioned example (Listing 1), the feature divides the texts into sentences, so the results are reported as the percentage over the number of sentences.
An advantage of UMUTextStats in comparison to other applications is that it allows operating simultaneously with different versions of the same text. Therefore, some dimensions can operate on a filtered version that makes it easier to look up terms in the dictionary, while the original version can be used to measure characteristics such as the percentage of words in capital letters.

- **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 suprasegmental feature concerning expressive lengthening, a linguistic device that consists in repeating some of the letters of a word for emphasis (Fersini et al., 2016).

- **Morphosyntax (MOR)**. Also known as grammar, it is the part of linguistics focused on morphology, which studies words and how they are composed, and syntax, which studies phrases and sentences and how they are related. Spanish is a highly inflected language. Inflections can denote multiple syntax and semantic meanings that can be used to track stylometric features in authorship attribution tasks. UMUTextStats classifies morphosyntactic features into: (1) PoS-based features, which include adverbs, adjectives, determiners, and pronouns, to name but a few; and (2) sub-word level, which includes features that capture subcomponents of words, such as stems and affixes. This includes features that capture grammatical gender and number of words.

- **Correction and style (CAS)**. This linguistic category covers linguistic and stylistic errors. Linguistic errors deviate from the valid rules of language. They include, for instance, misspellings or the wrong use of accentuation. Regarding stylistic errors, even though it is possible to understand a text containing them and the meaning conveyed by the author, the text may not sound natural to the recipient.

  UMUTextStats divides correction and style features into three major groups: orthographic, stylistic, and performance errors. Orthographic errors capture wrong uses in accentuation and misspellings. Besides, other writing mistakes, such as starting sentences in lowercase, are also identified as orthographic errors. Stylistic errors capture bad writing habits, such as starting sentences with cardinal numbers or repeating several sentences with the same word. Finally, performance errors detect if an author makes a mistake despite knowing well the rules of the language they are using.

  Besides, correction and style capture other common mistakes that include: (1) writing one word instead of two by mistake (*asique*, *sobretodo*, *portanto*); (2) writing two words instead of one (*a parte*, *sin fin*); (3) using nonexistent words (*deshaucio*, *inflacción*); (4) using an incorrect use of plural (*malostratos*); (5) poorly written Latinisms (*status quo*, *a grosso modo*); (6) skipping the accent mark in words where it is required (*aereo*, *duplex*); (7) use the accent mark in words that do not require it (*ávaro*), *fuí*, *pié*); and (8) incorrect and redundant expressions (*bajo mi punto de vista*, *detrás mío*).

- **Semantics (SEM)**. Semantics is the branch of linguistics that studies meaning out of context. Semantics can consider meaning at different units, such as words or sentences.

  The current version of UMUTextStats capture four linguistic categories, including (1) onomatopoeia, which is the formation of a word from a sound associated with what is named; (2) euphemism, which is a mild expression replacing one considered as too rude; (3) dysphemism, which is a derogatory expression used instead of a pleasant one; and (4) synecdoche, which is a figure of speech in which a part is used to represent the whole.

- **Pragmatics (PRA)**. It is the branch of linguistics that deals with how language is used and its context. UMUTextStats captures the use of figurative language, including hyperboles, several idiomatic expressions, verbal irony, understatements, metaphors, and similes. Besides, it captures rhetorical questions. Pragmatics also includes several discourse markers, used for structuring the conversation regarding connectors, reformers, argumentative clauses, and conversational bookmarks. Finally, it also includes several typical courtesy forms for greetings or con-

dolences.

- **Stylometry (STY)**. It is the automated study of linguistic style, mainly in written language. This category contains features concerning the length of the text, lexical diversity applying the standard type-token ratio (STTR) (Chipere et al., 2004). This metrics has been applied to calculate the lexical richness of texts as, for instance, in the clinical domain (López Hernández and Almela, 2021). Besides, we use different statistics to measure the number of words and syllables, number of sentences, number of words in uppercase, readability formulas or punctuation symbols.

- **Lexis (LEX)**. It refers to the total set of all possible words in a language, or a particular subset of words for a particular domain. It is worth mentioning that lexis provides a functional perspective, so an element of the lexis can be composed of multiple words, such as *New York City*.

  The current version of UMUTextStats includes lexicons of several general domains that capture the intention and the topic of the message. It includes lexicons of terms such as animals, weapons, jobs, crime, money, health, and ingesting, as well as abstract concepts such as achievement, risk, or cognitive processes.

- **Psycholinguistic processes (PLP)**. Psycholinguistics is the branch of linguistics concerning the usage of language from a cognitive point of view. It includes psychological processes involved in language comprehension, production, and first and second language acquisition. UMUTextStats captures LF based on lexicons, regarding emotions and positive, neutral, and negative attitudes.

- **Register (REG)**. The register helps to define the way in which the speakers and writers use the language under different circumstances. It is related to the selection of the words and other prosodic features related to tone, pitch and body language. UMUTextStats identifies features related to offensive language, informal speech, and the usage of learned words.

- **Social media (SOC)**. Although this category is not exclusive to linguistics, we added a set of features that capture the degree in which users of Twitter make use of specific terminology related to that network. They include the usage of hashtags, mentions, and hyperlinks. This category was added with the purpose of testing the extensibility of UMUTextStats and because we consider that it captures valuable features in certain domains such as cyberbullying or authorship attribution.

3.4. Output Resolver

This module is responsible for generating the output of the system. The current version of UMUTextStats has several outputs already predefined, including CSV, ARFF files (for the WEKA platform), HTML (for online visualisation), JSON, and Table format for outputting the results directly to the console.

UMUTextStats is ready to work from a terminal. However, it is also possible to use it from a web browser thanks to a graphic user interface.

4. Experimentation

UMUTextStats has been validated on different domains within the document classification task.

In Table 1 we include a summary of the results achieved in different shared tasks in which we evaluate the LF combined with embeddings.

Concerning SA (Eysenbach, 2002), the LF are applied to build an Ontology-guided Aspect-based Sentiment Analysis system from a dataset related to infectious diseases in Latin America (García-Díaz et al., 2020c). For this purpose, we build a custom ontology of the infectious disease domain to represent the aspects of the system. Next, we train a machine-learning classifier using ten-fold cross-validation and evaluating the LF and pretrained word embeddings. The neural networks evaluated include recurrent and convolutional neural networks C(NN). We observed that the LF in isolation achieved the best overall accuracy (55.3%). We also observed that all neural networks architectures except CNN achieved better accuracy when combined with the LF. Finally, we link the sentiment of each document with the aspects of the ontology by using a custom TF–IDF formula (Rodríguez-García et al., 2014), in which the sentiment of a document influences the concepts that appears both explicitly and implicitly in the text. In addition, we calculate the correlation between the LF and the sentiments. The results of this analysis is depicted in Figure 2 (top). We found a negative correlation between numerals and sentiments, and a positive correlation of colloquialism with positive documents.

Concerning HS, the LF are applied to conduct a misogyny detection system in (García-Díaz et al., 2021a), in which we also observed that LF also outperform features based on word and non-contextual sentence embeddings. We build several machine-learning classifier that evaluates all feature sets in isolation or combined. We obtained our best result with a model based on Support Vector Machine (SVM), achieving an accuracy of 85.175%. This result was achieved with the combination of the LF and sentence embeddings from fastText. Besides, we obtained the Information Gain for the 20 best LF. These features are depicted in Figure 2 (middle). We found positive correlations with the misogynous label with the usage of offensive language, grammatical gender, and several features concerning correction and style.

	Dataset	Notes	Score
SA	Infodemiology (García-Díaz et al., 2020c)	Ecuadorian Spanish, multiclass	55.3
	TASS 2020 (García-Díaz et al., 2020b)	Spanish, multi-class	35.8
HS	MeOffendes (Plaza-del Arco et al., 2021a)	Spanish, binary	87.82
	MisoCorpus (García-Díaz et al., 2021a)	Mexican, binary	90.52
	AMI 2018 (Fersini et al., 2018)	Spanish, binary	84.72
	HaterNET (Pereira-Kohatsu et al., 2019)	Spanish, binary	84.08
	HatEval 2019 (Basile et al., 2019)	Spanish, binary	77.27
	EXISTS 2021 (Rodríguez-Sánchez et al., 2021)	Spanish and English, binary	75.14
	EXISTS 2021 (Rodríguez-Sánchez et al., 2021)	Spanish and English, multi-class	53.62
AA	PoliCorpus 2020 (García-Díaz et al., 2022a)	Spanish, binary, gender	72.02
	PoliCorpus 2020 (García-Díaz et al., 2022a)	Spanish, multi-class, age range	46.69
	PoliCorpus 2020 (García-Díaz et al., 2022a)	Spanish, binary, ideology	98.04
	PoliCorpus 2020 (García-Díaz et al., 2022a)	Spanish, multi-class, ideology	91.05
	AI-SOCO 2020 (Fadel et al., 2020)	Multi-language, multi-class	91.16
FL	SatiCorpus 2021 (García-Díaz and Valencia-García, 2022)	Spanish, binary	97.32
	Satire (Barbieri et al., 2015)	Spanish, binary	90.00
	Satire (del Pilar Salas-Zárate et al., 2017)	Spanish, binary	95.63
	Satire (del Pilar Salas-Zárate et al., 2017)	Mexican, binary	92.84
	IrosVa 2019 (Ortega-Bueno et al., 2019)	Spanish, binary	70.94
	IrosVa 2019 (Ortega-Bueno et al., 2019)	Mexican, binary	66.40
	IrosVa 2019 (Ortega-Bueno et al., 2019)	Cuban, binary	66.34
	HaHackathon (Meaney et al., 2021)	English, binary	91.60
	HaHackathon (Meaney et al., 2021)	English, multiclass, controversial	46.50
	HaHa 2021 (Chiruzzo et al., 2021)	Spanish, binary	85.44
	HaHa 2021 (Chiruzzo et al., 2021)	Spanish, multiclass, mechanism	20.87
	HaHa 2021 (Chiruzzo et al., 2021)	Spanish, multiclass, target	32.25
EA	EmoEvalEs (Plaza-del Arco et al., 2021b)	Spanish, multiclass	66.84
	EmoEvalEs (Plaza-del Arco et al., 2021b)	Spanish, multiclass	66.84

Table 1: Evaluation of UMUTextStats organised by topic: SA (Sentiment-Analysis), HS (Hate-speech), AA (Author Analysis), FL (Figurative Language), and EA (Emotion Analysis). The score is based on F1-score

Nevertheless, the LF achieved more limited results in sexism identification (Garcıa-Dıaz et al., 2021c), in which we combined the LF with transformers in two ways: (1) EL and (2) KI. As this task was composed in the identification of sexist messages in Spanish and English, we rely only in the stylometric features for the English dataset.

In (García-Díaz et al., 2022b), two strategies for combining the LF with word and sentence embeddings are evaluated: KI and EL. These strategies are evaluated in four datasets: Spanish MisoCorpus 2020, HaterNET, AMI 2018, and HatEval 2019. The best performance is obtained with the KI strategy. Besides, regarding generic hate-speech identification, we evaluated the LF in MeOffendES (García-Dıaz, 2021), composed of texts written in European and Mexican Spanish containing hate-speech. Significantly, we achieved the second and fifth place in the binary classification tasks, and the first position in the subtasks with contextual features.

The LF are applied to Author Analysis. In (García-Díaz et al., 2022a), the LF evaluated (1) an author profiling task regarding psychographic and demographic traits, and (2) an authorship attribution task with tweets from Spanish politicians compiled in 2020. The results in this task were promising. In fact, LF in isolation outperformed BERT-based models in gender prediction. For the rest of the traits, we observed that adding LF to embedding-based features outperformed the results achieved separately. Besides, concerning interpretability of the results, we found that features related to lexis and morphosyntactic were more effective for conducting author profiling tasks whereas the stylometric features were more reliable in authorship attribution. The features of the multi-class political ideology are depicted in Figure 2 (bottom).

Besides, we evaluated the LF in other shared tasks proposed in IberEval 2021: one regarding Humour identification (Garcıa-Dıaz and Valencia-Garcıa, 2021a), in which we achieved the 1st position in the subtask of Funniness Score Prediction and the 3rd position for the subtask of target classification. The other task in which we evaluated the LF was related to emotion analysis (García-Díaz et al., 2021b), in which we achieved the 6th position.

It is worth noting that the current version of the LF can also be applied to other languages as some of the linguistic categories, like stylometric features,

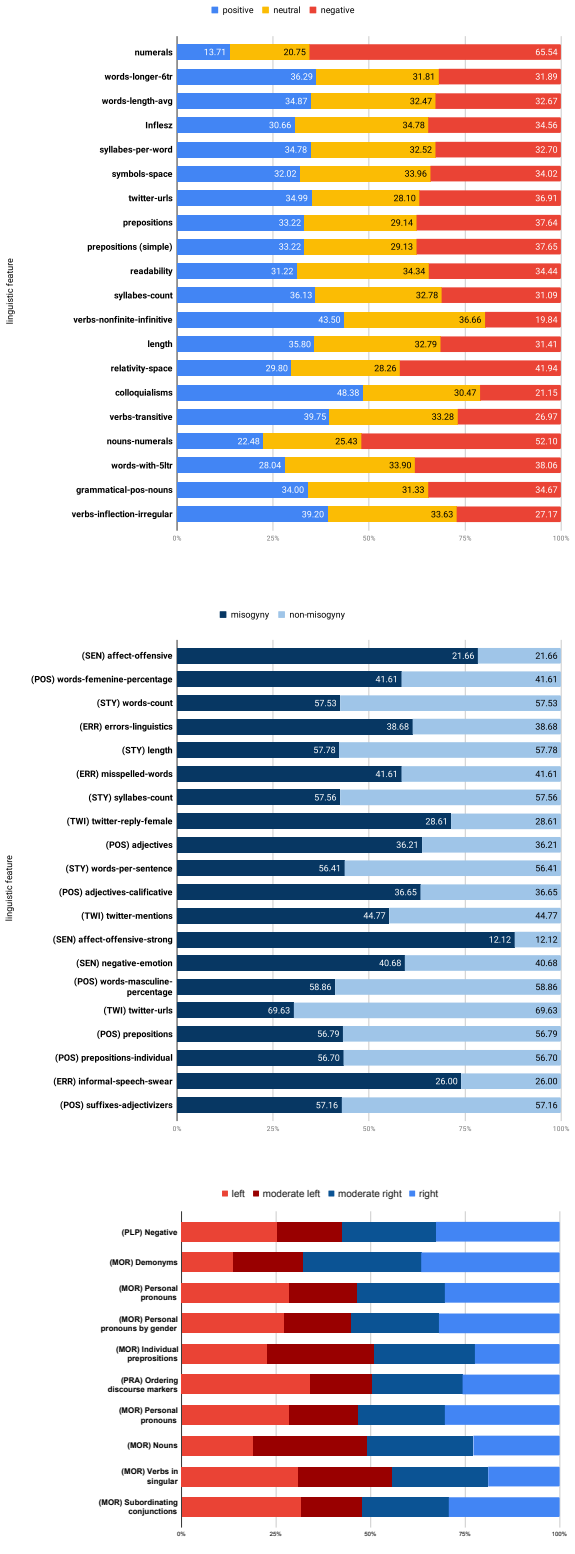

Figure 2: Information Gain of LF over Infodemiology (top) (García-Díaz et al., 2020c), over the Spanish MisoCorpus 2020 (middle) (García-Díaz et al., 2021a) and over the political multiclass with the Spanish PoliCorpus 2021 (bottom) (García-Díaz et al., 2022a)

are language independent. We participated in the AISOCO'2020 shared task (Fadel et al., 2020), focused on authorship identification of source-code. In this challenge, we proposed a system that mixes character n-grams with stylometric features that capture certain author traits. We achieved the 6th position (accuracy of 91.16%), outperforming baselines based on RoBERTa. Other languages in which the LF have been tested are English, Tamil, and Arabic. In English we have participated in HaHackathon 2021 (García-Díaz and Valencia-García, 2021b), concerning humor detection, achieving an F1-score of 91.60% concerning humour identification, and an F1-score of 57.22% in humour controversy detection. In Tamil we have participated in several shared tasks concerning equality, diversity, and inclusion. These tasks were focused on detecting depression signs, abusive comment detection, homophobic and transphobic comments, and emotion detection. In Arabic we have participated in a SemEval shared task concerning sarcasm detection.

5. Conclusions

In this paper we have described the development and the evaluation of the UMUTextStats tool, a tool for extracting linguistic features designed for the Spanish language. The evaluation is performed in several document classification tasks: Sentiment Analysis, Author Analysis, Satire identification, and Hate speech. There is a demo of the tool available[2]. All things considered, UMUTextStats still has some limitations. The main one is that dictionary-based dimensions consider words in isolation; so it is weak for polysemy and figurative language, especially in short texts such as the ones compiled from micro-blogging social networks. In order to solve this drawback, we are evaluating the reliability to obtain a small subset of LF per token, and train the models using Recurrent Neural Networks. In addition, we are adapting this tool to English and different Spanish dialects. Besides, we are extending and validating the hierarchical structure of the linguistic categories and features.

Acknowledgements

This work is part of the research project AI-InFunds (PDC2021-121112-I00), funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. This work is also part of the research project LaTe4PSP (PID2019-107652RB-I00), funded by MCIN/AEI/10.13039/501100011033. Besides, it was partially supported by Fundación Séneca -the Regional Agency for Science and Technology of Murcia (Spain)- through project 20963/PI/18. In addition, José Antonio García-Díaz is supported by Banco Santander and the University of Murcia through the Doctorado Industrial programme.

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

6. Bibliographical References

Almela, A., Valencia-García, R., and Cantos, P. (2013). Seeing through deception: A computational approach to deceit detection in spanish written communication. *Linguistic Evidence in Security, Law and Intelligence*, 1(1):3–12.

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

Barrio-Cantalejo, I., Simón-Lorda, P., Melguizo, M., Escalona, I., Marijuán, M., and Hernando, P. (2008). Validation of the inflesz scale to evaluate readability of texts aimed at the patient. In *Anales del sistema sanitario de Navarra*, volume 31, pages 135–152.

Basile, V., Bosco, C., Fersini, E., Debora, N., Patti, V., Pardo, F. M. R., Rosso, P., Sanguinetti, M., et al. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics.

Chipere, N., Malvern, D., and Richards, B. (2004). Using a corpus of children's writing to test a solution to the sample size problem affecting type-token ratios. *Corpora and language learners*, pages 139–147.

Chiruzzo, L., Castro, S., Góngora, S., Rosá, A., Meaney, J., and Mihalcea, R. (2021). Overview of haha at iberlef 2021: Detecting, rating and analyzing humor in spanish. *Procesamiento del Lenguaje Natural*, 67:257–268.

del Pilar Salas-Zárate, M., López-López, E., Valencia-García, R., Aussenac-Gilles, N., Almela, Á., and Alor-Hernández, G. (2014). A study on liwc categories for opinion mining in spanish reviews. *Journal of Information Science*, 40(6):749–760.

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

Eysenbach, G. (2002). Infodemiology: The epidemiology of (mis) information. *The American journal of medicine*, 113(9):763–765.

Fadel, A., Musleh, H., Tuffaha, I., Al-Ayyoub, M., Jararweh, Y., Benkhelifa, E., and Rosso, P. (2020). 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*.

Fersini, E., Messina, E., and Pozzi, F. A. (2016). Expressive signals in social media languages to improve polarity detection. *Information Processing & Management*, 52(1):20–35.

Fersini, E., Rosso, P., and Anzovino, M. (2018). Overview of the task on automatic misogyny identification at ibereval 2018. *Ibereval@ sepln*, 2150:214–228.

Garcıa-Dıaz, J. A. and Valencia-Garcıa, R. (2021a). 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*, volume 9.

García-Díaz, J. A. and Valencia-García, R. (2021b). 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)*, pages 1096–1101.

García-Díaz, J. A. and Valencia-García, R. (2022). Compilation and evaluation of the spanish saticorpus 2021 for satire identification using linguistic features and transformers. *Complex & Intelligent Systems*, pages 1–14.

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

García-Díaz, J. A., Almela, Á., Alcaraz-Mármol, G., and Valencia-García, R. (2020a). Umucorpusclassifier: Compilation and evaluation of linguistic corpus for natural language processing tasks. *Procesamiento del Lenguaje Natural*, 65:139–142.

García-Díaz, J. A., Almela, Á., and Valencia-García, R. (2020b). Umuteam at tass 2020: Combining linguistic features and machine-learning models for sentiment classification. In *IberLEF@ SEPLN*, pages 187–196.

García-Díaz, J. A., Cánovas-García, M., and Valencia-García, R. (2020c). Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in latin america. *Future Generation Computer Systems*, 112:641–657.

García-Díaz, J. A., Cánovas-García, M., Palacios, R. C., and Valencia-García, R. (2021a). Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings. *Future Gener. Comput. Syst.*, 114:506–518.

García-Díaz, J. A., Colomo-Palacios, R., and Valencia-Garcia, R. (2021b). Umuteam at emoevales 2021: Emosjon analysis for spanish based on explainable linguistic features and transformers. In *Proceedings of the Iberian Languages Evaluation Forum (Iber-LEF 2021), CEUR Workshop Proceedings, Málaga, Spain*, volume 9, pages 59–71.

Garcıa-Dıaz, J. A., Colomo-Palacios, R., and Valencia-Garcıa, R. (2021c). Umuteam at exist 2021: Sexist language identification based on linguistic features and transformers in spanish and english. In

Proceedings of the Iberian Languages Evaluation Forum (Iber-LEF 2021), CEUR Workshop Proceedings, Málaga, Spain, volume 9.

García-Díaz, J. A., Colomo-Palacios, R., and Valencia-García, R. (2022a). 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.

García-Díaz, J. A., Jiménez-Zafra, S. M., García-Cumbreras, M. A., and Valencia-García, R. (2022b). Evaluating feature combination strategies for hatespeech detection in spanish using linguistic features and transformers. *Complex & Intelligent Systems*, pages 1–22.

García-Dıaz, J. (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. CEUR-WS. org*.

López Hernández, J. and Almela, A. (2021). Detección automática de errores lingüísticos en textos clínicos: análisis de patrones de error en varias especialidades médicas. *Panace@*, 22(53):96.

López-López, E., del Pilar Salas-Zárate, M., Almela, Á., Rodríguez-García, M. Á., Valencia-García, R., and Alor-Hernández, G. (2014). Liwc-based sentiment analysis in spanish product reviews. In *Distributed Computing and Artificial Intelligence, 11th International Conference*, pages 379–386. Springer.

Martí Ferriol, J. L. (2016). Selection and validation of a measurement instrument for readability calculations in patient information leaflets for oncological patiens in spain.

Meaney, J., Wilson, S., Chiruzzo, L., Lopez, A., and Magdy, W. (2021). Semeval 2021 task 7: Hahackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 105–119.

Ortega-Bueno, R., Rangel, F., Hernández Farıas, D., Rosso, P., Montes-y Gómez, M., and Medina Pagola, J. E. (2019). Overview of the task on irony detection in spanish variants. In *Proceedings of the Iberian languages evaluation forum (IberLEF 2019), co-located with 34th conference of the Spanish Society for natural language processing (SEPLN 2019). CEUR-WS. org*, volume 2421, pages 229–256.

Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., and Camacho-Collados, M. (2019). Detecting and monitoring hate speech in twitter. *Sensors*, 19(21):4654.

Plaza-del Arco, F. M., Casavantes, M., Escalante, H. J., Martín Valdivia, M. T., Montejo Ráez, A., Montes y Gómez, M., Jarquín-Vásquez, H., and Villaseñor Pineda, L. (2021a). Overview of meoffendes at iberlef 2021: Offensive language detection in spanish variants.

Plaza-del Arco, F. M., Jiménez Zafra, S. M., Montejo Ráez, A., Molina González, M. D., Ureña López, L. A., and Martín Valdivia, M. T. (2021b). Overview of the emoevales task on emotion detection for spanish at iberlef 2021.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

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

Ramirez-Esparza, N., Chung, C. K., Kacewicz, E., and Pennebaker, J. W. (2008). The psychology of word use in depression forums in english and in spanish: Texting two text analytic approaches. In *ICWSM*.

Rodríguez-García, M. Á., Valencia-García, R., García-Sánchez, F., and Samper-Zapater, J. J. (2014). Ontology-based annotation and retrieval of services in the cloud. *Knowledge-based systems*, 56:15–25.

Rodríguez-Sánchez, F., Carrillo-de Albornoz, J., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., and Donoso, T. (2021). Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67:195–207.

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