IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 January 2023, accepted 7 February 2023, date of publication 10 February 2023, date of current version 15 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3244065

RESEARCH ARTICLE

Smart Analysis of Economics Sentiment in Spanish Based on Linguistic Features and Transformers

JOSÉ ANTONIO GARCÍA-DÍAZ, FRANCISCO GARCÍA-SÁNCHEZ, AND RAFAEL VALENCIA-GARCÍA

Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain

Corresponding author: Rafael Valencia-García (valencia@um.es)

This work is part of the research projects AIInFunds (PDC2021-121112-I00) and LT-SWM (TED2021-131167B-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/AEI/ 10.13039/501100011033) funded by MCIN/AEI/10.13039/501100011033. Besides, it was partially supported by the Seneca Foundation-the Regional Agency for Science and Technology of Murcia (Spain)-through project 20963/PI/18. In addition, Jose Antonio Garcia-Diaz is supported by Banco Santander and the University of Murcia through the Doctorado Industrial programme.

ABSTRACT Texts related to economics and finances are characterized by the use of words and expressions whose meaning (and the sentiments they convey) substantially depend on the context. This poses a major challenge to Natural Language Processing tasks in general, and Sentiment Analysis in particular. For low-resource languages such as Spanish, this situation becomes even more acute. Yet, the latest advancements in the field, including word embeddings and transformers, have allowed to boost the performance of Sentiment Analysis solutions. In this work we explore the impact of the combination of different feature sets in the accuracy of Sentiment Analysis in Spanish financial texts. For this, a corpus with 15,915 tweets has been compiled and manually annotated as either positive, negative, or neutral. Then, feature sets based on contextual and non-contextual embeddings along with linguistic features were evaluated both individually and combined. The best results, with a weighted F1-score of 73.15880%, were obtained with a combination of feature sets by means of knowledge integration.

INDEX TERMS Sentiment analysis, financial, transformers, feature engineering, deep learning.

I. INTRODUCTION

Recent trends in Natural Language Processing (NLP) and the development of pre-trained linguistic models based on transformers and attention mechanisms with large unannotated corpora are allowing to improve the accuracy of several NLP tasks such as named entity recognition (NER), automatic summarization, or sentiment analysis (SA) among others [1], [2]. Moreover, the development of multilingual assets and the conception of language specific linguistic models are allowing to improve the performance in low-resource languages such as Spanish [3].

Among specific domains from which information can be extracted, the language employed in the financial domain is particularly challenging. First, it contains words and expressions that are very specific to this domain. Moreover, it may contain terms such as anglicisms and acronyms for which pre-trained models may have few examples to learn from. Second, in some cases terms and phases are used in the financial domain with a meaning that differ from the true definition of the words. Expressions such as *Activo*[1] and *Pasivo*[2] could be easily misinterpreted as the adjectives *active* and *passive*, respectively. Non-contextual word embeddings are non-practical for dealing with those terms and they do not handle polysemy [4]. Third, the positive and negative polarity expressed in a document is not easily inferred using lexicon-based approaches. For example, as the increment of

The associate editor coordinating the review of this manuscript and approving it for publication was Ali Shariq Imran.

[1] In English: Financial asset.
[2] In English: Financial liability.

something could be seen as something positive in a general way, in the financial domain, the polarity could change with negative effects when it is risk premium or unemployment what increases.

In preliminary works, we have explored the reliability of applying non-contextual embeddings from Spanish pre-trained models for conducting SA to the financial domain [5]. For that purpose, we built a preliminary dataset with tweets written in Spanish concerned with the financial domain and manually labeled them as positive, negative, or neutral. This work represents a significant extension of our previous work in which we perform an in-depth evaluation of SA for the financial domain with the following major contributions: (1) we have enlarged and improved the dataset, which now contains 15,915 tweets labeled as positive, negative and neutral with tweets between 2017 and 2021, resulting in an increase of almost double; (2) we evaluate several multilingual and Spanish pre-trained models based on transformers, such as BETO [3], multilingual BERT [6], ALBETO [7], Distilled BETO [7], MarIA [8], BERTIN [9] and XLM [10]; (3) we evaluate the reliability of combining the feature sets by means of ensemble learning and knowledge integration; and, (4) we conduct an error analysis to understand the pros and cons of each feature set.

Although Spanish is one of the most widely spoken languages in the world[3] NLP resources are scarce. Therefore, the government of Spain, through the 'Plan for the Advancement of Language Technology' [11], is supporting the development of language models and annotated datasets to significantly improve NLP in Spanish. The novelty of our work is twofold. First, a new, manually-labeled sentiment dataset in Spanish is provided in the financial domain. Second, we analyze the performance of state-of-the-art models. Indeed, even though transformer-based architectures such as BERT and RoBERTa have been in the literature for some years (BERT was first published in 2018 and RoBERTa in 2019), these models were focused on English, and their multilingual counterparts do not offer satisfactory results when dealing with some NLP-tasks for Spanish texts as shown in our study. Therefore, given such limitations, it is necessary to validate new language models based on these architectures but centered on Spanish including BETO, ALBETO, MarIA and BERTIN. These models have been made available just recently and are still being validated in the Spanish language in different domains. As the financial domain is very specific, it is necessary to study whether these general language models provide good enough results for detecting sentiments in financial tweets written in Spanish. Moreover, just like in state-of-the-art researches, we analyze the mixture of different linguistic features with deep learning technologies based on transformers both to boost the accuracy of the results and to improve their interpretability.

The rest of the manuscript is organized as follows. Section II provides background information concerning SA, including attention mechanisms and transformer language models, and current approaches to SA in the financial domain. A detailed description of the corpus and its compilation process is presented in Section III. Then, in Section IV the system architecture is described along with a complete overview of all its components. The evaluation of the feature sets in isolation or combined and the error analysis conducted for the evaluation of the overall system is shown in Section V. Finally, conclusions and future work are put forward in Section VI.

[3]https://www.ethnologue.com/guides/ethnologue200

II. STATE OF THE ART

In this work, different feature sets based on linguistic features and transformer mechanisms are evaluated for dealing with SA in Spanish financial texts. In the last few years, researchers in the NLP field have explored different means to boost the overall classification success in SA [1]. The use of mechanisms of attention such as transformers along with the adoption of deep learning methods are proving successful for high resource languages such as English [2], but their impact on other, less resourced languages such as Spanish is yet to be validated. In this section, SA is introduced, and several Spanish and multilingual pre-trained contextual embeddings commonly used to improve the performance of NLP tools are enumerated. Then, various approaches to SA in the financial domain are discussed.

A. SENTIMENT ANALYSIS

The NLP task in which the subjective sentiment of a text is obtained is often referred to as Sentiment Analysis (SA), also known as, Opinion Mining [12]. Three levels of analysis depending on the degree of specificity can be distinguished, namely, document-level, sentence-level and aspect-level [1]. In document-level SA each document is labeled with a single sentiment, which is typically accurate on small documents providing general insights about the users' attitudes. However, when dealing with, for example, product reviews from online stores, different even contradictory opinions on the same product or service can be found. Under such circumstances, the sentence-level SA is more useful since a sentiment is calculated for each sentence in the document. The drawback here is that a manual revision is required to find out the topic each sentence is about. Aspect-level SA specifically deals with this issue by dividing texts into subtopics and assigning a sentiment to each one, thus becoming the most sophisticated approach for conducting SA [13].

The approaches for extracting the polarity of a text can be divided into the following three categories [14]: lexicon-based methods, machine learning-based methods and hybrid techniques that extend machine learning models with lexicon-based knowledge. The lexicon-based approach relies on words expressing positive or negative feelings to humans previously gathered and documented in a lexicon such as SentiWordNet [15]. Then, lexicon-based methods calculate the polarity scanning through the documents for these

keywords. Some linguistic phenomena, such as polysemy or ambiguity, can hamper the performance of this approach but their effects can be lessened with the use of domain specific lexicons [16], [17]. On the other hand, the machine learning-based approach consists of training a model to discriminate between positive, neutral, and negative texts. The use of supervised learning, in which the model is trained with labeled source data, often outperforms unsupervised and semi-supervised learning approaches, but depends on previously annotated examples, which is a time-consuming and subjective manual labor. In the last few years, researchers are exploring the use of deep neural networks, namely, deep learning, for SA [18]. Deep learning algorithms such as Long-Short Term Memory (LSTM), Convolutional Neural Networks (CNN), Gated Recurrent Units (GRU) and Recurrent Neural Networks (RNN) are some of the most common ones to accomplish this task [1].

The application of machine learning-based techniques in SA is contingent upon the extraction of meaningful features from the documents, commonly referred to as feature engineering. Three types of features are commonly distinguished, including statistical, linguistic, and contextual [19]. Traditional approaches generally use Bag-Of-Words (BoW), a statistical model that considers the frequency of the words within a vocabulary to represent a text in the form of a sparse vector. In line with this, the term frequency-inverse document frequency (TF–IDF) method puts the focus on words often occurring in one document but scarcely in the corpus. Differently, linguistic phenomena such as stylistic features characterize the linguistic approach. Part-Of-Speech (PoS) tagging is a popular process in which words are categorized with a part of speech depending on their context. Similarly, the Linguistic Inquiry and Word Count (LIWC) [20] tool is a text analysis program that calculates the percentage of words in a text that belong to one or more linguistic, psychological, and topical categories. LIWC has been used in opinion mining [21] and complex classification tasks, such as satire detection [22]. The main challenge of linguistic features is that they are not easily shared between languages and cultures and, consequently, the resulting models are largely language and cultural dependent. Finally, contextual features refer to information not explicitly expressed in the text, but available in its context such as author gender or time/date in which the text was written. The use of these contextual features is less discussed in the literature since their availability is not always ensured.

The state-of-the-art concerning modern feature engineering relies on the use of word embeddings [23]. Unsupervised generic tasks are typically used to learn these embeddings, which represent words or sentences using dense vectors with real numbers in which words with similar semantics are clustered together. However, polysemy constitutes a major challenge for traditional (non-contextual) embeddings since a word will be represented with the same vector regardless of its meaning in a sentence. Contextual word embeddings overcome this issue by taking into account the context of a word to generate the embeddings. In particular, the words that are next to a given word are considered when producing its representation. Word embeddings constitute a statistical approach typically used in deep learning models. One of the many advantages of word embeddings is that they can be used with different neural networks architectures such as CNN and RNN, that exploit the spatial and temporal dimension of the text, respectively. However, while CNN models have been successful to reduce the dimensionality of the feature space and to extract meaningful features from texts in SA applications, attention mechanisms help tackle another major issue in this field, namely, the need for focusing on the important parts of the contextual information of texts [2]. Attention mechanisms are used to focus on the important parts of the context by assigning different weights [24]. Transformers constituted a major improvement boosting the speed with which models that use attention can be trained. Originally introduced in [25], the Transformer is based solely on attention mechanisms without the need for RNN or CNN in the encoder-decoder configuration. BERT [26], a large pre-trained Transformer network, has become one of the most popular language models across various NLP tasks. RoBERTa [27] constitutes an optimization of BERT pre-training procedure outperforming BERT models in almost all NLP tasks. Then, the research focus now is on building domain specific language models that better capture the domain features. An example in the financial domain is FinBERT ([28] or [29]), a language model pre-trained on large-scale financial corpora.

Similarly, BERT and other transformer models are being adapted to specific languages other than English. BETO [3], for example, is an initiative to allow the use of BERT pre-trained models for Spanish NLP tasks. Another example is ALBETO [7], which is a version of ALBERT [30] (which, in turn, is a lightweight version of BERT) that has been pre-trained only with documents written in Spanish. Likewise, MarIA [8] is based on RoBERTa and has been trained with text gathered from the National Library of Spain. Finally, BERTIN is also based on RoBERTa and has been trained with the Spanish split of the mC4 dataset [31].

The Spanish Society for Natural Language Processing (SEPLN) organizes every year since 2012 the 'Workshop on Semantic Analysis at SEPLN' (TASS)[4], which focuses on SA for the Spanish language. The original task of TASS is the evaluation of polarity classification systems of tweets written in Spanish and different variants (Spain, Peru, Costa Rica, Chile, Uruguay, and Mexico). In Section V-D we compare the performance of the approach proposed in this work against the proposal that achieved the best results in the last edition of the shared task, that took place in 2020.

To conclude this SA overview, in a recent review Osorio Angel et al. [32] studied the latest advances in SA for the Spanish language. While the pipeline and the techniques used at each step (information extraction, preprocessing,

[4]http://tass.sepln.org/

feature extraction, sentiment classification and evaluation) are comparable to those used for any other language, the authors highlight the ever-growing number of linguistic resources (e.g., lexicon or corpus) explicitly developed for the Spanish language. In terms of performance, deep learning models achieved the best results.

B. SENTIMENT ANALYSIS IN THE FINANCIAL DOMAIN

Product reputation, customer experience, market research and stock price prediction are some the areas within the financial domain in which SA has proven useful [33]. In this section, the most recent results in this field and the main approaches are discussed.

Going back to 2012, the authors in [34] described an experiment using SA for market analysis, in which a strong correlation between market and public sentiments was discovered. Similarly, in [35] the authors gathered tweets from big cap technological stocks and used SA to check volatility, trading volume and stock prices. Again, a high correlation between the stock prices and the extracted sentiments was found. Later, in 2014 Uhr et al. [36] introduced a method for sentiment analysis in financial markets combining word associations through the 'Concept for the Imitation of the Mental Ability of Word Association' (CIMAWA) and lexical resources. They evaluated the evolution of stock prices as compared to the sentiment measures calculated from a news corpus with 918,427 finance-related German documents. Sentiment values were obtained at document, sentence, and window size levels. The authors concluded that sentiment analysis in large time scales can assist in financial market-related decision making and risk management. In [37] the authors compare various deep learning models (LSTM, doc2vec and CNN) in predicting the semantic polarity of contributions to a financial social network such as StockTwits. They used Chi-square, analysis of variance (ANOVA) and mutual information as feature selection methods and BoW-based logistic regression (LR) as a baseline. From StockTwits, the authors collected the 140-characters messages posted by users in the first six months of 2015 and determined that CNN was the most effective model with a 90.9% accuracy.

More recently, the authors in [38] used the sentiments extracted from news articles along with other key indicators to build a predictive model from a fundamental analysis perspective. The target in [39] is markedly different. It constitutes a study on error patterns of some sentiment analysis methods commonly used in the financial domain. In their experiments, they make use of two datasets for the finance domain, namely, the Yelp dataset and the StockTwits Sentiment (StockSen) dataset and investigate eight representative models belonging to one out of the three explored approaches: lexicon-based (OpinionLex, SenticNet, and L&M), machine learning-based (Support Vector Machines (SVM) and fastText), and deep learning NLP models (BiLSTM, S-LSTM, and BERT). Six common causes for financial sentiment analysis errors are identified, i.e., irrealis mood, rhetoric, dependent opinion, unspecified aspects, unrecognized words, and external reference. But the latest contributions in this area put the focus on the use of SA for stock forecasting applying a variety of deep learning-based approaches. For example, in [40] the authors employ a CNN model for classifying the investors' sentiments from Chinese posts in a stock forum and then propose a LSTM-based system that take into account the sentiment analysis results along with further technical indicators to predict stock prices. In particular, the CNN-based sentiment analysis model outperformed other compared approaches (LR, SVM, RNN and LSTM) when dealing with the 880,000 posts gathered from Eastmoney.com reaching a F-measure value of 0.8482. In a very similar fashion, Shi et al. [41] present an individual stock movement prediction algorithm in which the sentiment polarity of Chinese comments posted in a financial online community are used in conjunction with the trade values of the stock in the previous five days. While SVM and LR models are employed in the stock movement predictor, CNN- and RNN-based algorithms are utilized in the sentiment classifier (with LR as baseline). For word embedding, a 300-dimensional size was found optimal using word2vec. The best sentiment classification results are obtained with GRU, reaching a F-measure value of 0.83. In contrast, the approach suggested in [42] relies on a number of proxy variables (i.e., the number of newly opened A-share accounts, the market turnover rate, the number of monthly IPO, discount of closed-end funds and first-day return of IPO) to build an investor sentiment index using the PCA (Principal Component Analysis) method. Then, the dynamic relationship between stock market returns, investor sentiment, and volatility are studied, finding an asymmetric influence of investor sentiment on the stock market, with a lower impact in the bearish stock market than in the bullish stock market.

The number of works concerned with SA for Spanish in the financial domain is scarcer. In [43], the authors claim that sentiment indicators about a given entity are traditionally treated as silos that cannot be combined and propose a Linked Data-based approach in which the sentiment information from different communities can be integrated. The proposed system gathers tweets from Twitter and builds a corpus in which only tweets related to financial institutions are kept. Then, each tweet is represented in the form of triplets which are annotated using SentiWordNet; the arithmetic mean of all registered values are used to quantify the tweet sentiment. The authors in [44] present a domain specific lexicon called FSAL focused on the financial domain and experimentally prove that combining different machine learning techniques for SA with this domain specific lexicon results in better classification performance than using a generic lexicon. In their experiment, they make use of three machine learning algorithms, namely, Naive Bayes (NB), Random Forest (RF) and SVM, to process a corpus with 500 tweets in Spanish. In a recent report published by the *Banco de España* [45], a Spanish dictionary of words with positive, negative or

TABLE 1. Corpus statistics.

Label	Train	Val	Test	Total
Positive	2505	835	836	4176
Neutral	4669	1556	1557	7782
Negative	2374	791	792	3957
Total	9548	3182	3185	15915

neutral connotation in the context of financial stability is described, which is then used to create sentiment indices from financial texts. To calculate the sentiment index, the number of words in the text with a negative connotation and the number of words with a positive connotation are considered. The texts from both the *Banco de España*'s Financial Stability Reports and press reports are used to evaluate the robustness of the built dictionary to estimate the sentiment of texts in this domain. Finally, in a previous work [5], we explored the reliability of applying non-contextual pre-trained embeddings to the financial domain. Specifically, the embeddings were trained from two main sources, namely, the Spanish Unannotated Corpora (SUC) and the Spanish Billion Word Corpus and Embeddings (SBWC). The embeddings were trained using different methods, such as Word2Vec, fastText and GloVe. These resources were used to train a deep learning classifier to extract the polarity of Spanish tweets from the financial domain in an earlier version of the datasets presented in this work. The best accuracy achieved was 58.036% using a GRU with fastText and SUC.

A comprehensive evaluation of the most significant approaches to SA in finance is provided in [46]. According to their results, NLP transformers present the highest performance and even their distilled versions (e.g., DistilledBERT) obtain promising results in text classification tasks.

III. CORPUS

For compiling the dataset, we relied on the UMUCorpus-Classifier tool [47], developed by our research group. This tool crawls data from Twitter, a social network in which users can send and receive micro-blogging posts of less than 280 characters. We extracted tweets from popular Spanish economists and news sites focused on the financial domain. The tweets are between November 2017 and October 2021 and they have been manually labeled in different stages and by different annotators. Preliminary versions of this dataset were published in [5], [48].

In a nutshell, the labeling process can be described as follows: each annotator should identify a tweet with the following sentiments: *very-positive*, *positive*, *neutral*, *negative*, *very-negative*, *out-of-domain*, and *do-not-know-do-not-answer*. It is worth noting that a tweet can be labeled more than once. The average number of ratings of a tweet is 2.5342. In addition, the annotators achieve a inter-agreement score based on the Krippendorff's alpha [49] of 0.63058 of a total of 32,028 annotations. Table 1 shows the corpus statistics.

FIGURE 1. An example of a Tweet of the corpus.

An example of a tweet from the corpus is shown in Figure 1.[5]

The next step of the corpus compilation consisted in discarding those tweets labeled as *out-of-domain*. Next, tweets labeled as *positive* were merged with those labeled as *very-positive*, and so were *negative* tweets with *very-negative* ones. In case that a tweet received contradictory labels from different annotators, it was considered as *neutral*. At the end of this process, we obtained 15915 tweets: 4176 positive tweets, 7782 neutral tweets, and 3957 negative tweets. The corpus is available at the following URL.[6] This file contains the Twitter's IDs of the tweets, the label, and the split (training, validation, and testing). We do not include the tweets due to Twitter guidelines,[7] which advises not to share the text, so users preserve the rights to delete their content on the Internet.

IV. SYSTEM ARCHITECTURE

To validate the feature sets and the deep learning architectures a system was built whose architecture is depicted Figure 2.

In brief, the system works as follows. First, to clean the dataset a text preprocessing phase is applied (see Section IV-A). Second, the dataset is divided into training, validation, and testing subsets in a 60-20-20 ratio (see Section III). Third, a feature extraction stage is conducted to obtain the linguistic features and the embedding-based features (see Section IV-B). Fourth, we evaluate three strategies for evaluating the features: (1) single feature evaluation, (2) knowledge integration, and (2) ensemble learning. During this stage, an hyper-parameter optimization phase is performed to evaluate different deep learning architectures (see Section IV-D). In the end, the test dataset is used to evaluate the best deep learning models for each feature integration strategy.

A. TEXT PREPROCESSING

The Text Preprocessing module generates different versions of the texts: (1) original, (2) normalized, and (2) normalized with lowercase. The original version is used to extract

[5]In English: 'The development of 5G could bring Spain benefits of 14.6 billion euros in 2021'.

[6]http://pln.inf.um.es/corpora/economics/economics-2021.rar

[7]https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases

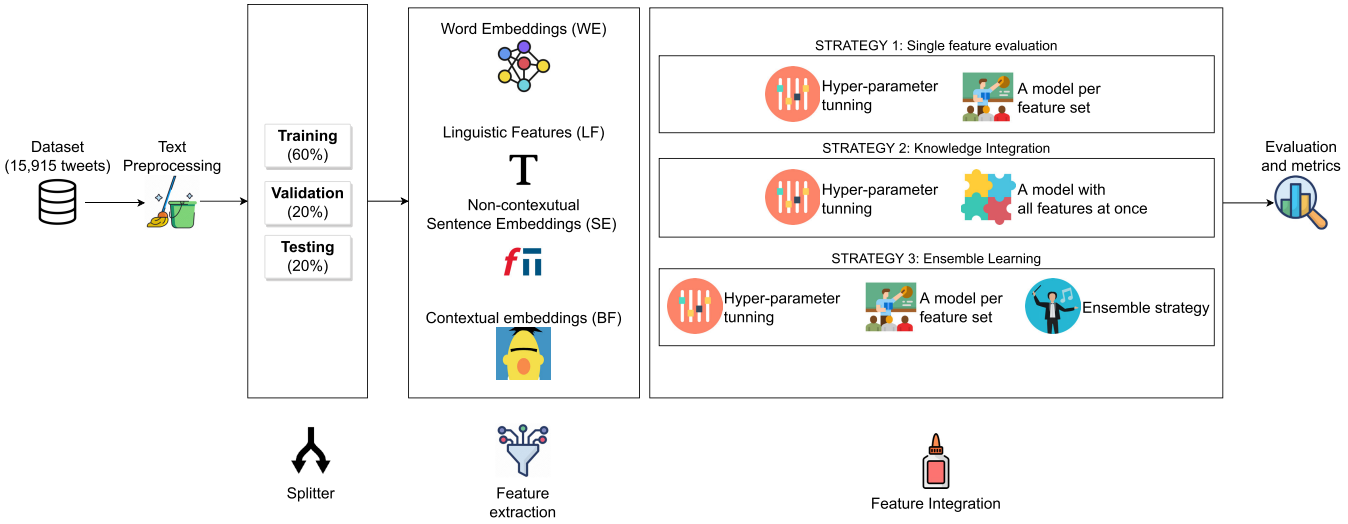


FIGURE 2. System architecture.

certain linguistic features related to correction and style. The normalized version is used as source to extract PoS features. To do this, hyperlinks, hashtags, or mentions among other social media jargon are removed. In addition, all digits and percentages are replaced with a fixed token because we do not want deep learning classifiers to learn specific quantities. We also strip expressive lengthening, by removing repetitions of the same letter that occurs more than twice. In addition, we fix misspellings with the ASPELL tool.[8] Finally, the normalized version in lowercase form is used to extract the tokens of the embeddings based features.

B. FEATURE EXTRACTION

Following the spirit of our previous works [50], we evaluated several feature sets for conducting opinion mining in the financial domain. Specifically, linguistic features along with contextual and non-contextual pre-trained embeddings were evaluated.

For the extraction of the **linguistic features (LF)**, we rely on UMUTextStats [19], [51], which was developed by our research group inspired by LIWC [20]. This tool captures 365 linguistic features organized as follows:

- **Correction and style of the writing communication (COR)**. It distinguishes orthographic errors, including statistics concerning misspelled words, stylistics and bad performance errors, such as sentences starting with numbers or with the same word, or common errors and redundant expressions.
- **Phonetics (PHO)**. It captures expressive lengthening, that is, the intentional elongation of some letters with an emphasizing purpose.
- **Morphosyntax (MOR)**. It captures how words are composed, including grammatical gender, number, and a great variety of affixes, including nominals, adjectivizers, verbalizers, adverbializers, augmentative, diminutives, or derogatory suffixes. It also captures and organizes features according to their PoS category (e.g., verbs, nouns, adjectives, etc.). For this, we mix Stanza [52] with lexicons that capture fine grained categories.
- **Semantics (SEM)**. It includes onomatopoeia, euphemism, dysphemism, and synecdoche.
- **Pragmatics (PRA)**. It captures figurative language devices, including understatements, rhetorical questions, hyperbole, idiomatic expressions, verbal irony, metaphors, or similes among others.
- **Stylometry (STY)**. It captures punctuation symbols, corpus statistics, and other metrics related to the number of words, syllables, or sentences.
- **Lexical (LEC)**. It captures the topics in the text analyzing both abstract and general topics.
- **Psycho-linguistic processes (PLI)**. Emojis and lexicons related to emotions and sentiments are considered in this category.
- **Register (REG)**. It captures the presence of informal or cultured language along with topics related to offensive speech.
- **Social media jargon (SOC)**. Features associated to the speaker's mastery of social media jargon are captured in this category.

As for the **embeddings**, both contextual and non-contextual word and sentence embeddings were explored as follows:

- **Non-contextual word embeddings (WE)**. Pre-trained models based on word2vec [53], fastText [54], and gloVe [55] are evaluated. As mentioned above, word embeddings enable the exploration of specific types of neural network architectures, such as CNN and RNN, which can take advantage of the space and temporal dimensions of language, respectively. In particular, CNN

[8] http://aspell.net/

can generate high-order features by clustering multi-word terms whose meaning differs from the one it can be obtained taking each word separately (e.g., *New* and *York*). On the other hand, RNN leverage the temporal dimension by considering the order of the words. In this work, two bidirectional RNN based on LSTM and GRU have been evaluated, namely, BiLSTM and BiGRU.

- **Non-contextual sentence embeddings (SE)**. These are extracted from FastText [56], whose Spanish model is trained from CommonCrawl and Wikipedia [57].
- **Contextual word embeddings (BF)**. Different pre-trained transformer-based models are evaluated. These models can classified as BERT-based models and RoBERTa-based models. The key difference between these two architectures is that in RoBERTa, the masking is performed during training time whereas in BERT the masking is performed at the beginning of the training. The architectures based on BERT are: (i) BETO, the Spanish version of BERT [3]; two novel lightweight versions: (ii) ALBETO and (iii) Distilled BETO [7], and (iv) multilingual BERT (mBERT). The architectures based on RoBERTa are (i) MarIA [8], (ii) BERTIN [9] and (iii) XLM [10]. The HuggingFace transformers library was used to fine tune the models with the corpus. It is important to bear in mind that given that these kinds of embeddings are very time consuming, they are difficult to combine with other feature sets. Therefore, the fixed representation of the `[CLS]` token is extracted as suggested in [58]. Then, this representation is used to combine the contextual embeddings more easily with the rest of the feature sets. Preliminary results indicate that the precision, recall, and accuracy of both contextual word and sentence embeddings are similar with this and other datasets.

Once feature sets are extracted, a feature normalization and selection step is conducted. A MinMax scaler is applied first to the linguistic features as they contain features in different scales with raw counts and percentages. Then, Information Gain is applied to select the best features, discarding those that belong to the last quartile.

C. FEATURE INTEGRATION: KNOWLEDGE INTEGRATION AND ENSEMBLE LEARNING

This part of the pipeline is responsible for the integration of the feature sets to build more robust solutions. Particularly, three strategies are evaluated as follows: single feature evaluation, knowledge integration and ensemble learning.

The first strategy, single feature evaluation, does not combine any feature set. Several models are trained for each feature set separately using hyper-parameter tuning and the best model is selected using the custom validation split. In this sense, we use this strategy as baseline to compare the other two strategies.

The second strategy for combining the features is known as knowledge integration, which consists in the integration of different feature sets within the same neural network. For this, a multi-input deep learning network is trained from scratch. The idea is that the network learns during training how to combine the strengths of each feature set. The network architecture design followed in this work is to connect each feature set into a different stack of hidden layers. Then, the output of each layer is concatenated and connected to a new stack of hidden layers that are connected to the final output layer. As it is not clear which is the best network architecture for this task, we conducted the same hyper-parameter tuning stage (see Section IV-D) to get the best model, so different number of neurons, hidden layers, and activation functions are evaluated.

The third strategy, ensemble learning, consists of combining the predictions of several estimators (that are run separately) to build a more robust estimator [59]. In this work we checked four averaging methods to combine the predictions of each feature set. These averaging methods are as follows: (1) mode, which outputs the label with a majority vote of each estimator; (2) weighted mode, which is similar to the hard voting strategy, but the contribution of each model is weighted according to the performance of each model with the validation set; (3) highest probability, which consists of observing the probability of each label and model and select the higher; and (4) average probabilities, which averages the probabilities output by each model.

D. HYPER-PARAMETER EVALUATION

The next step in our pipeline is to conduct the hyper-optimization stage. The main objective of this phase is to find out what are the best parameters for the neural networks. Ir order to do this, for each feature set (in isolation and in combination) we trained several models and ranked them by weighted F1-score, considering the label distribution. For each feature set, we tested different neural network architectures. For LF, SE, and BF we relied on multilayer perceptron (MLP) as these features do not contain sequence information such as text. In case of word embeddings, we evaluated CNN and bidirectional RNN based on LSTM (BiLSTM) and GRU (BiGRU). Moreover, we evaluated randomly its shape, composed by the number of hidden layers and the number of neurons in each layer. The communication between the layers is made by several activation functions. We also included a dropout mechanism to avoid overfitting. The dropout is configured in a ratio of 10%, 20%, 30% or not using dropout at all. In addition, two more parameters were evaluated, namely, the batch size and the learning rate. All neural networks made use of a time-based learning rate scheduler. The best hyper-parameters for each feature set and their combinations can be seen in Table 2.

From Table 2, it can be observed that the majority of models that achieve the best results make use of shallow neural networks composed of one or two hidden layers. The number of neurons per layer is large in case of sentence embeddings (256 for SE, 512 for BF) but smaller for LF (8) and WE (3). Concerning the dropout, all the

TABLE 2. Hyper-parameters for each feature set and their combinations.

	architecture	shape	layers	neurons	dropout	lr	activation
baseline	dense	rhombus	3	48	0.1	0.001	sigmoid
LF	dense	brick	1	8	0.1	0.010	sigmoid
SE	dense	brick	1	256	0.1	0.001	relu
WE	dense	brick	2	3	0.1	0.001	relu
BF	dense	brick	1	512	0.3	0.010	sigmoid
LF-SE	dense	brick	1	48	0.3	0.001	relu
LF-WE	dense	brick	1	256	False	0.001	sigmoid
LF-BF	dense	brick	1	16	0.2	0.001	sigmoid
SE-WE	dense	brick	1	128	False	0.001	tanh
SE-BF	dense	brick	2	3	0.1	0.010	sigmoid
WE-BF	dense	brick	1	3	0.3	0.001	relu
LF-SE-we	dense	brick	1	3	0.1	0.010	tanh
LF-SE-BF	dense	brick	2	3	0.3	0.010	tanh
LF-WE-BF	dense	lfunnel	5	48	False	0.001	sigmoid
SE-WE-BF	dense	lfunnel	5	48	False	0.001	sigmoid
LF-SE-WE-BF	dense	brick	1	3	False	0.001	sigmoid

features in isolation achieved better results with smaller dropout (10% for LF, SE, WE; 30% with BF). When the features are combined in the same neural network, some combinations achieved better results without dropout, as is the case of LF-WE, SE-WE, LF-WE-BF, SE-WE-BF, and the combination of all feature sets (LF-SE-WE-BF).

V. RESULTS AND DISCUSSION

To compare the results achieved by the rest of the feature sets, we set a baseline based on a BoW. This statistical model analyzes the frequency of the words in the documents. The drawbacks of BoW models are various. First, they consider words without context, so they do not take into account linguistic phenomena such as polysemy. Second, they create over-fitted models due to the vocabulary size, thus suffering the *curse of dimensionality*. Moreover, in agglutinative languages this problem is aggravated, since there are a large number of words that can be obtained from the same root. These problems are partially solved with the usage of bigrams and trigrams or the use of char-grams, as they are able to capture lexical information, including the use of punctuation symbols and morphological information, such as prefixes or suffixes. In addition, char-grams are more robust against grammatical errors, as misspelled words and their correct versions share common char-grams.

Specifically, we extracted the TF–IDF of unigrams, bigrams, and trigrams as well as character n-grams of length 3, 4, and 5, applying sub-linear scaling, from the lowercase version of the tweets. Next, we reduced the length of the features by applying Latent Semantic Analysis (LSA) to 100 components. We have selected this method because these features are easy to extract, and they provide good results in several classification tasks. However, these features have some shortcomings. First, they are features that have lost the ordering and the meaning of the words. Second, they suffer from the *curse of dimensionality*, as its size depends on the vocabulary size. This handicap has been partially overcome by applying LSA.

Since we address an imbalanced classification problem, we evaluated the performance of the deep learning models with the weighted-average F1-score (see Equation 1), that is the harmonic mean between Precision (see Equation 2), and Recall (see Equation 3) of each class but weighted by the number of instances in each class. In addition, the Accuracy measure (see Equation 4) has been also considered. In the equations, TP stands for *true positive*, TN stands for *true negative*, FP stands for *false positive* and FN stands for *false negative*.

$$F1 = 2 * (Precision * Recall)/(Precision + Recall) \tag{1}$$

$$Precision = TP/(TP + FP) \tag{2}$$

$$Recall = TP/(TP + FN) \tag{3}$$

$$Accuracy = TP + TN/(TP + TN + FP + FN) \tag{4}$$

A. RESULTS OF THE FEATURE SETS IN ISOLATION

Prior to the experimentation stage, we evaluated different multilingual and Spanish pre-trained models based on Transformers. The idea was to determine which one is the most accurate in order to use the best pre-trained model to combine the features with the rest. Accordingly, Table 3 shows the results achieved by the pre-trained models based on Transformers. The evaluated models are: (i) BETO [3], a Spanish BERT trained with the Spanish Unannotated Corpora; (ii) ALBETO [7], a version of ALBERT, which is a lightweight version of BERT, pre-trained only with Spanish documents; (iii) Distilled BETO [7], trained using distillation techniques to transfer the weights of BETO to a new model with less layers and complexity; (iv) MarIA [8], based on RoBERTa and trained with web crawlings from the National Library of Spain; (v) BERTIN [9], which is another model based on RoBERTa, trained with the Spanish split of the mC4 dataset; (vi) multilingual BERT [6], a BERT

TABLE 3. Pre-trained Transformer-based models.

Pre-trained Model	Precision	Recall	F1-score	Accuracy
BETO	70.950 70	70.989 01	70.962 22	70.989 01
ALBETO	70.355 51	69.165 34	69.691 05	71.177 39
Distilled BETO	68.472 22	68.083 05	68.261 92	69.733 12
MarIA	72.550 74	**72.464 68**	**72.188 47**	72.464 68
BERTIN	**71.561 13**	71.557 04	71.493 31	**72.621 66**
mBERT	70.077 21	70.047 10	69.957 32	70.047 10
XLM	71.093 51	69.786 28	70.356 38	71.805 34

model trained with documents written in 104 languages; and (vii) XLM [10], a multilingual version of RoBERTa, trained with data filtered from CommonCrawl from 100 different languages. It can be observed that the RoBERTa architecture achieves better results than BERT. The two best scores are achieved by MarIA (72.188%) and BERTIN (71.493%). Besides, XLM outperforms multilingual BERT (70.356% vs 69.957%). It can also be observed that the lightweight versions of BETO, ALBETO and Distilled BETO, reach similar results to this first one. In fact, ALBETO achieves a better accuracy than BETO. When comparing the results achieved with transformers trained from single language datasets versus those trained from multilingual datasets, it can be seen that there is a slight advantage of the models trained only with Spanish, which suggest that it is preferable to obtain specific pre-trained models for the target language rather than to use multilingual variants.

In Table 4 the results of the feature sets in isolation are presented. As it can be observed, the baseline of TF–IDF of word and character n-grams with LSA achieved a weighted F1-score of 55.97398%. The rest of the feature sets improve this baseline. The best result is achieved with BF, reaching a weighted F1-score of 68.554336%. This result outperforms the weighted F1-score of LF (58.46046%), SE (65.18025%), and WE (65.14914%). It can be highlighted that the embeddings (i.e., SE, WE, BF) are more effective than LF for sentiment classification in the financial domain, both in terms of precision and recall. This fact suggests that the lexical features and what is said is more important than the linguistic features that better capture the tone and style of the authors.

We can observe that all feature sets achieve similar results regarding the overall precision and recall of the system. However, considering the precision and recall of the features individually (not shown in the table but available with the source code[9]), one can see that BF is the feature set in which these measures are more similar regardless the label (i.e., either positive, negative, or neutral). The precision and recall among all classes achieved with SE are also similar. Yet, WE achieve more precision with the positive class but less recall (precision of 64.03162%, recall of 58.13397%). The same is observed with the neutral class (precision of 74.44134%, recall of 68.46500%) but the opposite behavior with the negative class (precision of 52.61570%, recall of 66.035353%). In LF, however, the system achieves better recall than precision with the positive class (precision of 50.40984%, recall of 58.85167%) but better precision than recall with the neutral class (precision of 69.12568%, recall of 64.99680%) and the negative class (precision of 47.51678%, recall of 44.69697%).

[9]https://github.com/NLP-UMUTeam/Smart-Analysis-of-Economics-in-Spanish

B. RESULTS OF KNOWLEDGE INTEGRATION

The results of the knowledge integration experiment are provided in Table 5. The table is organized in combinations of two, three and four feature sets.

Concerning the feature sets combined in pairs, we can observe that the results are generally superior to the ones achieved individually. These results suggests that the feature sets are complementary. However, the results do not always improve the ones achieved by the best individual model. For example, the combination of LF and SE achieved a weighted F1-score of 64.17839%. This result improves largely the results achieved by LF (58.46046%) but it is worse than SE (65.18025%). The best overall result is obtained with the combination of three feature sets: LF, WE, and BF, achieving a weighted F1-score of 73.15880%. These results improve the combination of all features (72.98100%).

The degradation of the results when adding non-contextual sentence embeddings (SE) to the LF-WE-BF combination might be due to merging embeddings that cannot handle polysemy in the same way.

C. RESULTS OF ENSEMBLE LEARNING

The results of the ensemble learning experiment are shown in Table 6. The best weighted F1-score is achieved by applying the average probabilities strategy with an F1-score of 72.48612%. This result is slightly worse than the best result achieved by the knowledge integration strategy (73.15880% of F1-score with the combination LF-WE-BF). The result achieved by the highest probability strategy is similar (72.35689%) to the one obtained with average probabilities. It draws our attention the good results achieved by the highest probability strategy since other experiments conducted by our research group showed that this strategy usually achieves good precision but limited recall in binary classification experiments. Concerning the mode and the weighted mode, the results are even lower with an F1-score of 69.42306% and 71.53359%, respectively. These results indicate that the weighted mode is more accurate as it considers the performance of each model.

Summing up, the best result in terms of F1-score is obtained with the knowledge integration strategy, combining LF, WE and BF feature sets. The next best result (without considering other knowledge integration-based combinations) is achieved by using ensemble learning with the average probabilities strategy followed by the RoBERTa-based model and the BF isolated model. These results can be partially due to the fact that knowledge integration strategies can learn patterns that occur when certain linguistic features and certain

TABLE 4. Results of the feature sets in isolation.

Feature set	Architecture	Precision	Recall	F1-score	Accuracy
baseline	MLP	55.566 10	56.954 47	55.973 98	56.954 47
LF	MLP	58.839 75	58.335 95	58.460 46	58.335 95
SE	MLP	65.410 45	65.023 55	65.180 25	65.023 55
WE	MLP	66.281 71	65.149 14	65.149 14	65.149 14
BF	MLP	**72.086 97**	**72.150 71**	**68.544 36**	**71.949 31**

TABLE 5. Results of the knowledge integration experiment.

Feature set	Architecture	Precision	Recall	F1-score	Accuracy
LF-SE	MLP	64.104 03	64.270 02	64.178 39	64.270 02
LF-WE	MLP	65.754 76	66.091 05	65.685 84	66.091 05
LF-BF	MLP	72.302 37	72.276 30	72.288 12	72.276 30
SE-WE	MLP	67.891 09	66.970 17	67.178 85	66.970 17
SE-BF	MLP	69.350 20	69.042 39	69.112 97	69.042 39
WE-BF	MLP	71.863 86	71.930 93	71.862 52	71.930 93
LF-SE-WE	MLP	66.998 06	66.562 01	66.725 73	66.562 01
LF-SE-BF	MLP	72.190 28	72.182 10	72.156 87	72.182 10
LF-WE-BF	MLP	**73.182 64**	**73.186 81**	**73.158 80**	**73.186 81**
SE-WE-BF	MLP	72.975 11	72.715 86	72.408 56	72.715 86
LF-SE-WE-BF	MLP	72.971 98	73.029 83	72.981 00	73.029 83

TABLE 6. Results of the ensemble learning experiment of LF, SE, WE, and BF.

Strategy	Precision	Recall	F1-score	Accuracy
highest probability	72.499 68	72.558 87	72.356 89	72.558 87
average probabilities	**72.537 11**	**72.684 46**	**72.486 12**	**72.684 46**
weighted mode	71.511 66	71.648 35	71.533 59	71.648 35
mode	70.066 06	69.419 15	69.423 06	69.419 15

embeddings take place at the same time. At this point, it is also worth highlighting that it is easier to adopt the ensemble learning strategy than the knowledge integration strategy as it is not necessary to train and perform a hyper-parameter selector over a new model with a large number of parameters.

D. COMPARISON WITH RELATED WORK

The TASS 2020 dataset[10] has been selected to compare the reliability of our methods with existing datasets. This shared task proposed a sentiment classification task at three levels with tweets written in different variants of Spanish. The variants are Spain (ES), Costa Rica (CR), Peru (PE), Uruguay (UR), and Mexico (MX). The dataset is evaluated from both a monolingual and a multi-variant perspectives.

Table 7 reports the results of applying our best model (i.e., a combination of Linguistic features (LF), contextual sentence embeddings from RoBERTa (BF), and non-contextual word embeddings (WE) using knowledge integration) to the dataset provided by the TASS 2020 shared task. As reported in [60], the teams ELiRF-UPV, for the ES, CR PE and MX variants, and Palomino-Ochoa, for the UR variant and the multi-variant challenge, achieved the best results.

We can observe that our proposal outperforms the best macro F1-score in ES (69.2% vs 67.1%) and MX (64.7% vs 63.4%) but it is more limited for CR (63.8% vs 64.6%), PE (61.7% vs 63.6%) and UR (65.0% vs 66.9%). In the case of PE, our system outperforms the macro precision (69.1% vs 67.2%) and macro recall (60.7% vs 60.3%).

It is worth mentioning that for this TASS 2020 dataset different ensemble learning strategies outperform the results achieved with knowledge integration. We obtain a macro F1-score of 71% in ES with the highest probability strategy, a macro F1-score of 65.8% in CR with the weighted mode strategy, a macro F1-score of 63.9% in PE with the highest probability strategy, a macro F1-score of 67.1% in UR with highest probability strategy, and a macro F1-score of 65.77% in MX with the average probabilities strategy. However, none of the ensemble learning strategies outperform the knowledge integration result with the multi-variant dataset, getting a best macro F1-score of 46% with the average probabilities strategy as compared with the 46.3% achieved with the LF-WE-BF knowledge integration-based combination.

From these results we can conclude that our model is competitive in terms of macro precision, recall, and F1-score in different Spanish dialects and that the knowledge integration strategy achieves better performance with texts written in different Spanish variants whereas ensemble learning strategies are better suited for one specific Spanish variant.

E. INTERPRETABILITY OF THE FEATURES

To discover the contribution of each linguistic feature to the subjective polarity of the document, we calculated the Information Gain [61] of each feature. The top-10 metrics and how they contribute to each label are shown in Figure 3. We can observe that there are several features related to morphology, such as the percentage of nouns, adverbs, and

[10]http://tass.sepln.org/2020/?page_id=74

TABLE 7. Comparison with TASS-2020. In the left, we report the macro averaged precision, recall and F1-score of our model based on LF, WE, and BF trained with knowledge integration. In the right, we report the same metrics for the winner of the TASS-2020 best result.

Dataset	Precision	Recall	F1-score	Precision	Recall	F1-score
ES	69.2	69.2	69.2	67.3	67.0	67.1
CR	64.1	63.7	63.8	64.6	64.7	64.6
PE	69.1	60.7	61.7	67.2	60.3	63.6
UR	65.5	65.0	65.0	67.1	66.7	66.9
MX	65.2	64.5	64.7	63.7	63.3	63.4
ALL	47.2	46.8	46.3	48.7	51.0	49.8

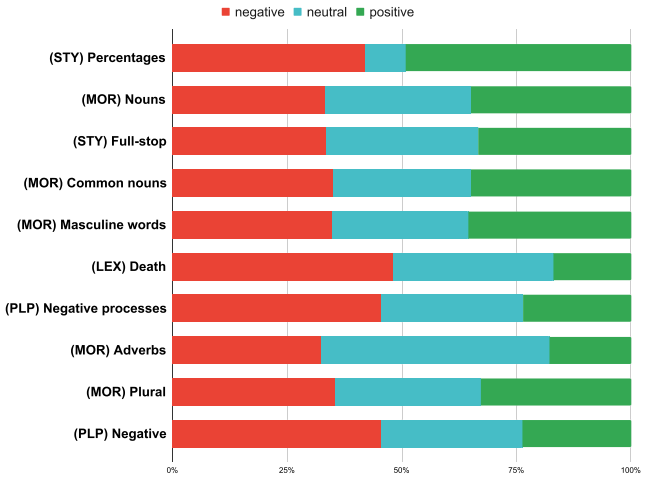


FIGURE 3. Information gain.

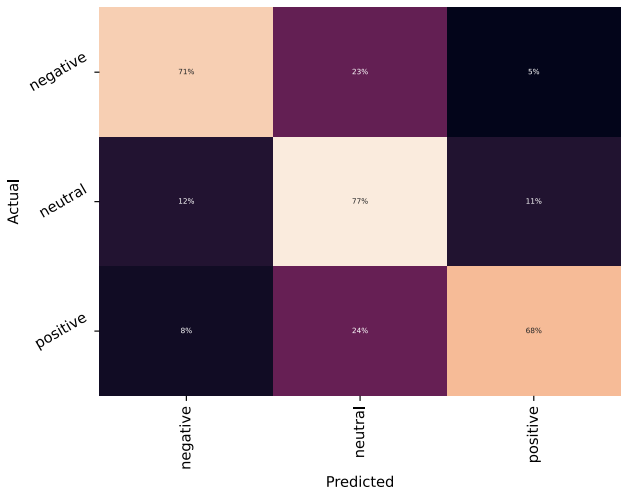


FIGURE 4. Confusion matrix of the combination of LF, WE and BF in the same neural network.

words in plural. Out of these, only the percentage of adverbs has more relevance to neutral documents. The most relevant stylometric linguistic feature is the usage of the percentage symbol, which appears very frequently in positive and negative texts, but rarely seen in neutral ones. This finding suggests that it is easy to infer the subjective polarity of tweets that report objective facts and statistics. Another linguistic category that constitutes good LF is Lexical, with words and expressions related to death, that are more common in negative documents, and negative processes from the psycho-linguistic processes category.

F. ERROR ANALYSIS

For conducting the error analysis, we made use of the neural model that provided best weighted F1-score over the test split, that consisted in the combination of LF, WE, and BF in the same neural network. Prior to error analysis, to check in what cases the model gives wrong predictions the confusion matrix has been plotted (see Figure 4). Particularly, taking into account this confusion matrix it can be observed that the model does not make many relevant wrong classifications (i.e., mismatching positive and negative documents), and the focus can be set on the relationship between neutral documents and their prediction as either positive or negative, and vice versa. We observe that the model labeled wrongly a 23% of negative and a 24% of positive documents as neutral. In addition, the ratio of wrong classifications of neutral documents is not skewed for negative nor positive labels, being the percentage of wrong classifications of 12% and 11%, respectively.

Next, to assess the overall performance of our best model, we compared the predictions with the ones obtained by the baseline model based on character and word n-grams. We observed that there are no instances that are correctly classified by the baseline model but not by our best model. This finding suggests that our best model completely outperforms the baseline. It is worth noting that the instructions for replicating these results can be found in the code repository.

Finally, we selected all the cases that were wrongly classified by our best model and, specifically, we focused on those predictions in which the neural model outputs a probability of the opposite label with a chance larger than 50% (i.e., either the ground truth is positive or negative but with a prediction probability of the opposite label equal or larger than 50%). However, under these specifications it is not possible to find any wrong prediction. Consequently, we changed the threshold from 50% to 45% and found three instances, which are listed in Table 8.

We can observe that there were only three wrong predictions in all the test dataset. Out of these, there is only an instance of a positive document that is wrongly classified as neutral by a slightly superior probability (0.47847% chance of being neutral vs 0.45145% probability of being positive).

TABLE 8. Error Analysis. We include the text, the ground truth (label), and the probabilities that the case belong to the negative, neutral and positive class as produced by the model.

text	label	p. negative	p. neutral	p. positive
Candriam apuesta por las acciones europeas y emergentes para [NUMERO] y descarta la recesión	positive	0.07008	0.47847	0.45145
El precio de la vivienda de segunda mano sube un [NUMERO]% en agosto, según idealista	negative	0.45024	0.05209	0.49767
La fiebre por las juntas telemáticas en el Ibex congela a los minoritarios en bolsa	negative	0.45923	0.50011	0.04066

This document informs that a company is committed to European and emerging stocks, ruling out recession. There are also two negative documents wrongly classified, one as positive and the other as neutral. The first document deals with the raising of the price of second-hand housing. This document is assigned a probability of 0.45024% of being negative and a 0.49767% of being positive, with a negligible probability of being neutral. In this sense, we highlight one of the problems with the financial domain in which a document could be positive to some actors in the financial market, but also at the same time negative for others. Finally, the last document is about how minority shareholders on the IBEX 35[11] stock exchange have encountered difficulties to be able to exercise their rights due to online meetings.

VI. CONCLUSION AND FURTHER WORK

In this paper we have explored the reliability of applying different feature sets based on contextual and non-contextual embeddings with linguistic features to improve Sentiment Analysis in Spanish for a challenging context such as the financial domain. Our results indicate that the combination of feature sets by means of knowledge integration provides the best results with a weighted F1-score of 73.15880%.

Additionally, the results obtained by other strategies, such as ensemble learning, are quite similar in terms of performance with the added benefits of (i) facilitating the combination of the feature sets and (ii) being more easily trainable. From our experiment it can be also noted that the usage of contextual word embeddings based on attention mechanisms represents a qualitative leap when it comes to improving the accuracy of the models. Further, the usage of general linguistic features provides limited results regarding sentiment analysis although they improve a baseline based on n-grams with LSA. Then again, the usage of linguistic features improves the accuracy when combined with transformers.

The benefits of knowledge integration over ensemble learning is that the neural network can learn from different feature sets at once. So, with knowledge integration the network can learn what features are more relevant for each document and how to combine them resulting in more general solutions. In ensemble learning, however, the fact that certain feature sets achieve, generally, better performance than others is ignored, except in some texts. For instance, linguistic features can make a significant difference with transformers in cases when some linguistic clues that are hard to obtain with transformers can be guessed. This is the case of expressive lengthening, for instance. Therefore, the disparate performance of each of the used feature set limits the performance of ensemble learning. This behavior is not observed in all the strategies, because it is more sensitive in strategies such as average probabilities or the mode of predictions than in weighted mode or highest probability. In this sense, we argue that ensemble learning is more effective when all feature sets have similar performance or when several models are trained with the same feature sets but varying hyper-parameters, such as the seed or the learning rate.

As future work, we are compiling a large corpus from Spanish financial newspapers to retrain BERT and RoBERTa based models, applying masked language modeling, and then fine-tune the model for sequence classification and perform an error analysis to check whether the accuracy of the system is improved or not. Another promising research line is the application of Semantic Web and ontologies to guide an Aspect-based Sentiment Analysis [19]. In this sense, we can explore how sentiments are transmitted through entities and their relationships and discover new types of products in which to invest in the short or medium term. Moreover, it would be possible to model different types of customers and their preferences by extracting demographic and psychographic information and build recommender systems [62] that can assist companies and individuals to carefully choose their investments based on their preferences.

[11]The Iberian Index (IBEX 35) is the Spain's principal stock exchange index.

REFERENCES

[1] A. Ligthart, C. Catal, and B. Tekinerdogan, "Systematic reviews in sentiment analysis: A tertiary study," *Artif. Intell. Rev.*, vol. 54, no. 7, pp. 4997–5053, Oct. 2021.

[2] M. E. Basiri, S. Nemati, M. Abdar, E. Cambria, and U. R. Acharya, "ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis," *Future Gener. Comput. Syst.*, vol. 115, pp. 279–294, Feb. 2021.

[3] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez, "Spanish pre-trained BERT model and evaluation data," in *Proc. ICLR*, 2020, pp. 1–10.

[4] A. Miaschi and F. Dell'Orletta, "Contextual and non-contextual word embeddings: An in-depth linguistic investigation," in *Proc. 5th Workshop Represent. Learn. (NLP)*, 2020, pp. 110–119.

[5] J. A. Garcia-Diaz, O. Apolinario-Arzube, and R. Valencia-Garcia, "Evaluating pre-trained word embeddings and neural network architectures for sentiment analysis in Spanish financial tweets," in *Proc. Mex. Int. Conf. Artif. Intell.* Cham, Switzerland: Springer, 2020, pp. 167–178.

[6] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[7] J. Cañete, S. Donoso, F. Bravo-Marquez, A. Carvallo, and V. Araujo, "ALBETO and DistilBETO: Lightweight Spanish language models," 2022, *arXiv:2204.09145*.

[8] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pamies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, and M. Villegas, "Maria: Spanish language models," *Procesamiento Del Lenguaje Natural*, vol. 68, pp. 39–60, Jan. 2022.

[9] J. D. L. Rosa, E. G. Ponferrada, P. Villegas, P. G. D. P. Salas, M. Romero, and M. Grandury, "BERTin: Efficient pre-training of a Spanish language model using perplexity sampling," *Procesamiento del Lenguaje Natural*, vol. 68, pp. 13–23, Jul. 2022.

[10] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzman, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 2019, *arXiv:1911.02116*.

[11] M. O. E. Affairs, D. Transformation, and G. O. Spain. (Oct. 2015). *Plan for the Advancement of Language Technology*. Accessed: Nov. 7, 2022. [Online]. Available: https://plantl.mineco.gob.es/tecnologias-lenguaje/PTL/Bibliotecaimpulsotecnologiaslenguaje/Detalle%20del%20Plan/Plan-Advancement-Language-Technology.pdf

[12] R. K. Bakshi, N. Kaur, R. Kaur, and G. Kaur, "Opinion mining and sentiment analysis," in *Proc. 3rd Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, 2016, pp. 452–455.

[13] Y. Tian, L. Yang, Y. Sun, and D. Liu, "Cross-domain end-to-end aspect-based sentiment analysis with domain-dependent embeddings," *Complexity*, vol. 2021, pp. 1–11, Mar. 2021.

[14] O. Kolchyna, T. P. T. Souza, C. P. Treleaven, and T. Aste, "Twitter sentiment analysis: Lexicon method, machine learning method and their combination," 2015, *arXiv:1507.00955*.

[15] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*, vol. 10, 2010, pp. 2200–2204.

[16] J. M. Ruiz-Martinez, R. Valencia-Garcia, and F. Garcia-Sanchez, "Semantic-based sentiment analysis in financial news," in *Proc. 1st Int. Workshop Finance Econ. Semantic Web*, 2012, pp. 38–51.

[17] M. Du, X. Li, and L. Luo, "A training-optimization-based method for constructing domain-specific sentiment lexicon," *Complexity*, vol. 2021, pp. 1–11, Feb. 2021.

[18] M. Vicari and M. Gaspari, "Analysis of news sentiments using natural language processing and deep learning," *AI Soc.*, vol. 36, no. 3, pp. 931–937, Sep. 2021.

[19] J. A. García-Díaz, M. Cánovas-García, and R. Valencia-García, "Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in Latin America," *Future Gener. Comput. Syst.*, vol. 112, pp. 641–657, Nov. 2020.

[20] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *J. Lang. Social Psychol.*, vol. 29, no. 1, pp. 24–54, Mar. 2010.

[21] M. del Pilar Salas-Zárate, E. López-López, R. Valencia-García, N. Aussenac-Gilles, Á. Almela, and G. Alor-Hernández, "A study on LIWC categories for opinion mining in Spanish reviews," *J. Inf. Sci.*, vol. 40, no. 6, pp. 749–760, Dec. 2014.

[22] M. D. P. Salas-Zárate, M. A. Paredes-Valverde, M. Á. Rodriguez-García, R. Valencia-García, and G. Alor-Hernández, "Automatic detection of satire in Twitter: A psycholinguistic-based approach," *Knowl.-Based Syst.*, vol. 128, pp. 20–33, Jul. 2017.

[23] C. Periñán-Pascual, "Measuring associational thinking through word embeddings," *Artif. Intell. Rev.*, vol. 55, no. 3, pp. 2065–2102, Mar. 2022.

[24] S. Kardakis, I. Perikos, F. Grivokostopoulou, and I. Hatzilygeroudis, "Examining attention mechanisms in deep learning models for sentiment analysis," *Appl. Sci.*, vol. 11, no. 9, p. 3883, Apr. 2021.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, N. A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, I. Guyon, U. V. Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds. Long Beach, CA, USA, 2017, pp. 5998–6008.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.

[27] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[28] D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models," 2019, *arXiv:1908.10063*.

[29] Z. Liu, D. Huang, K. Huang, Z. Li, and J. Zhao, "FinBERT: A pre-trained financial language representation model for financial text mining," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 4513–4519.

[30] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," in *Proc. 8th Int. Conf. Learn. Represent.*, Addis Ababa, Ethiopia, Apr. 2020, pp. 1–17.

[31] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "MT5: A massively multilingual pre-trained text-to-text transformer," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2021, pp. 483–498.

[32] S. O. Angel, A. P. P. Negrón, and A. Espinoza-Valdez, "Systematic literature review of sentiment analysis in the Spanish language," *Data Technol. Appl.*, vol. 55, no. 4, pp. 461–479, Aug. 2021.

[33] L. J. Gandía and D. Huguet, "Análisis textual y del sentimiento en contabilidad: Textual analysis and sentiment analysis in accounting," *Revista de Contabilidad-Spanish Accounting Rev.*, vol. 24, no. 2, pp. 168–183, Jul. 2021.

[34] A. Mittal and A. Goel, "Stock prediction using Twitter sentiment analysis," Dept. Manage. Sci. Eng., Standford Univ., Standford, CA, USA, Tech. Rep., CS229, 15, 2012.

[35] T. Rao and S. Srivastava, "Analyzing stock market movements using Twitter sentiment analysis," in *Proc. ASONAM*, 2012, pp. 119–123.

[36] P. Uhr, J. Zenkert, and M. Fathi, "Sentiment analysis in financial markets a framework to utilize the human ability of word association for analyzing stock market news reports," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2014, pp. 912–917.

[37] S. Sohangir, D. Wang, A. Pomeranets, and T. M. Khoshgoftaar, "Big data: Deep learning for financial sentiment analysis," *J. Big Data*, vol. 5, no. 1, pp. 1–25, Jan. 2018.

[38] A. Picasso, S. Merello, Y. K. Ma, L. Oneto, and E. Cambria, "Technical analysis and sentiment embeddings for market trend prediction," *Exp. Syst. Appl.*, vol. 135, pp. 60–70, Nov. 2019.

[39] F. Xing, L. Malandri, Y. Zhang, and E. Cambria, "Financial sentiment analysis: An investigation into common mistakes and silver bullets," in *Proc. 28th Int. Conf. Comput. Linguistics*, Barcelona, Spain, 2020, pp. 978–987.

[40] N. Jing, Z. Wu, and H. Wang, "A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction," *Exp. Syst. Appl.*, vol. 178, Sep. 2021, Art. no. 115019.

[41] Y. Shi, Y. Zheng, K. Guo, and X. Ren, "Stock movement prediction with sentiment analysis based on deep learning networks," *Concurrency Comput., Pract. Exper.*, vol. 33, no. 6, p. e6076, Mar. 2021.

[42] J. Hu, Y. Sui, and F. Ma, "The measurement method of investor sentiment and its relationship with stock market," *Comput. Intell. Neurosci.*, vol. 2021, pp. 1–11, Mar. 2021.

[43] J. Fernando Sánchez-Rada, M. Torres, C. A. Iglesias, R. Maestre, and E. Peinado, "A linked data approach to sentiment and emotion analysis of Twitter in the financial domain," in *Proc. 2nd Int. Workshop Semantic Web Enterprise Adoption Best Practice 2nd Int. Workshop Finance Econ. Semantic Web Co-Located 11th Eur. Semantic Web Conf.*, vol. 1240, A. García-Crespo, J. M. Gómez-Berbís, M. Radzimski, J. L. Sánchez-Cervantes, S. Coppens, K. Hammar, M. Knuth, M. Neumann, D. Ritze, and M. V. Sande, Eds. Anissaras, Greece, 2014, pp. 1–12.

[44] J. P. Braña, M. J. A. Litterio, and A. Fernández, "Fsal: Lexicón financiero de sentimiento en español rioplatense diseñado para 'bolsas y mercados argentinos' (BYMA)," *Revista Abierta de Informática Aplicada*, vol. 2, pp. 5–22, Sep. 2021.

[45] I. M. Bernal and C. G. Pedraz, "Sentiment analysis of the Spanish financial stability report," Banco de España, Madrid, España, Tech. Rep. 2011, 2020.

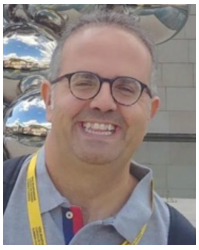
[46] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev, and D. Trajanov, "Evaluation of sentiment analysis in finance: From lexicons to transformers," *IEEE Access*, vol. 8, pp. 131662–131682, 2020.

[47] J. A. García-Diaz, A. Lmela, G. Alcaraz-Mármol, and R. Valencia-García, "Umucorpusclassifier: Compilation and evaluation of linguistic corpus for natural language processing tasks," *Procesamiento del Lenguaje Natural*, vol. 65, pp. 139–142, Jan. 2020.

[48] J. A. García-Díaz, M. P. Salas-Zárate, M. L. Hernández-Alcaraz, R. Valencia-García, and J. M. Gomez-Berbís, "Machine learning based sentiment analysis on Spanish financial tweets," in *Proc. World Conf. Inf. Syst. Technol.* Cham, Switzerland: Springer, 2018, pp. 305–311.
[49] K. Krippendorff, "Reliability in content analysis: Some common misconceptions and recommendations," *Hum. Commun. Res.*, vol. 30, no. 3, pp. 411–433, Jul. 2004.
[50] J. A. García-Díaz and R. Valencia-García, "Compilation and evaluation of the Spanish SatiCorpus 2021 for satire identification using linguistic features and transformers," *Complex Intell. Syst.*, vol. 8, no. 2, pp. 1723–1736, Apr. 2022.
[51] J. A. García-Díaz, M. Cánovas-García, R. Colomo-Palacios, and R. Valencia-García, "Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings," *Future Gener. Comput. Syst.*, vol. 114, pp. 506–518, Jan. 2021.
[52] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python natural language processing toolkit for many human languages," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, 2020, pp. 101–108.
[53] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. 1st Int. Conf. Learn. Represent. (ICLR)*, Y. Bengio and Y. LeCun, Eds. Scottsdale, AZ, USA, 2013, pp. 1–12.
[54] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," in *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*, 2018, pp. 1–4.
[55] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[56] K. Krasnowska-Kieraś and A. Wróblewska, "Empirical linguistic study of sentence embeddings," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 5729–5739.
[57] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," in *Proc. 11th Int. Conf. Lang. Resour. Eval. (LREC)*, Miyazaki, Japan, May 2018, pp. 1–5.
[58] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 1–11.
[59] O. Sagi and L. Rokach, "Ensemble learning: A survey," *WIREs Data Mining Knowl. Discovery*, vol. 8, no. 4, p. e1249, Jul. 2018.
[60] M. G. Vega, M. C. Díaz-Galiano, M. G. Cumbreras, F. M. P. D. Arco, A. Montejo-Ráez, S. M. J. Zafra, E. M. Cámara, C. A. Aguilar, M. A. S. Cabezudo, L. Chiruzzo, and D. Moctezuma, "Overview of TASS 2020: Introducing emotion detection," in *Proc. Iberian Lang. Eval. Forum Co-Located 36th Conf. Spanish Soc. Natural Lang. Process. (SEPLN)*, vol. 2664, M. A. G. Cumbreras, J. Gonzalo, E. M. Cámara, R. Martínez-Unanue, P. Rosso, S. M. J. Zafra, J. A. O. Zambrano, A. Miranda, J. P. Zamorano, Y. Gutiérrez, A. Rosá, M. M. Gómez, and M. G. Vega, Eds. Malaga, Spain, 2020, pp. 163–170.
[61] N. Patel and S. Upadhyay, "Study of various decision tree pruning methods with their empirical comparison in WEKA," *Int. J. Comput. Appl.*, vol. 60, no. 12, pp. 20–25, Dec. 2012.
[62] F. García-Sánchez, R. Colomo-Palacios, and R. Valencia-García, "A social-semantic recommender system for advertisements," *Inf. Process. Manag.*, vol. 57, no. 2, Mar. 2020, Art. no. 102153.

JOSÉ ANTONIO GARCÍA-DÍAZ received the B.E., M.Sc., and Ph.D. degrees in computer science from the University of Murcia, Espinardo, Spain. He is currently a Post-Doctoral with the Department of Informatics and Systems, University of Murcia. He has participated in more than ten research projects. He has published over 40 articles in journals, conferences, and book chapters. His research interests include natural language processing technologies focused on automatic text classification, sentiment and emotion analysis and hate speech detection. In addition, he has been part of the organizing committee for shared-tasks at international conferences such as IberLEF or EvalITA.

FRANCISCO GARCÍA-SÁNCHEZ received the B.E., M.Sc., and Ph.D. degrees in computer science from the University of Murcia, Murcia, Spain, in 2003, 2005, and 2007, respectively. From May 2012 to January 2017, he worked as the Vice Dean of External Relations with the Faculty of Computer Science. Formerly, he was a Ph.D. Assistant Professor with the Escola Superior Tècnica d'Enginyeria (ETSE), University of Valencia. He is currently an Associate Professor with the Department of Informatics and Systems, University of Murcia. He has taken part both as a principal investigator and a researcher in several research projects related to the application of semantic web technologies to real world challenges. He has published over 70 articles in journals, conferences, and book chapters. He has conducted a number of research stays in some of the most prestigious, semantic web- and AI-concerned research institutes around the world, such as the Semantic Technology Institute (STI, formerly the Digital Enterprise Research Institute, DERI), the Centre for Information Technology Research (CITR), and the Stanford Research Institute (SRI). His research interests include semantic web-based applications, including ontologies, linked data and knowledge graphs, natural language processing, semantic service-oriented architectures, social semantic web, and the application of AI technologies (including machine and deep learning).



RAFAEL VALENCIA-GARCÍA received the B.E., M.Sc., and Ph.D. degrees in computer science from the University of Murcia, Espinardo, Spain. He is currently a Full Professor with the Department of Informatics and Systems, University of Murcia. He has participated in more than 35 research projects. He has published over 150 articles in journals, conferences, and book chapters, 50 of them in JCR-indexed journals. He is the author or coauthor of several books. His research interests include natural language processing, semantic web, and recommender systems. He has been a Guest Editor of five JCR-indexed journals, such as *CSI*, *IJSEKE*, *JRPIT*, *JUCS*, and *SCP*.