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

José Antonio García-Díaz¹ · Rafael Valencia-García¹

Received: 23 July 2021 / Accepted: 17 December 2021 / Published online: 5 January 2022
© The Author(s) 2022

Abstract

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

Keywords Satire identification · Text classification · Feature engineering · Linguistic corpus · Natural language processing

Introduction

From ages, the spirit of satire have seek to make, through humor and mockery, a constructive criticism of society. Satire has the ability to make the weak strong and the strong weak. There are no relevant people who are immune to the sharpen comments of satyrs, that have not hesitated to parody the figure of kings, pontiffs or politicians. This way, satire reveals the vices of the society and raises people's awareness concerning rooted social conventions that are taboo and difficult to report by means other than humor. Apart from traditional satirical press, the rise of social media have allowed individuals to impersonate and parody relevant people with the same critical sense that satire possesses. It is important, however, to not to confuse satire with fake news and propaganda, which objective is to confuse people and influence political opinion with half-truths or hoaxes.

According to the Poe's law,¹ popular on the Internet culture, parody and extreme views are impossible to discern without clear indicators of the writer's intentions. This fact occurs very frequently when satirical, real news, and hoaxes are shared in social media by users or even by other news media who get confused and take satirical news and hoaxes as real. If satire is hard to identify for humans, their automatic identification is a even more challenging task. Satyrs rely on figurative language, in which words move away from their conventional meaning. Specifically, satire makes use of sarcasm and irony to play with words to hide criticism. Figurative language hinders Information Retrieval (IR) tasks that may misinterpret satire as truthful information. Moreover, there is a need to improve resources related to languages other than English in order to improve automatic satire identification.

As far as our knowledge goes, automatic satire identification in Spanish was first addressed by [3] and then by [21], in which the authors proposed, respectively, novel datasets containing tweets from media sites and build automatic classifiers by means of linguistic and statistical features together with traditional machine learning approaches such as Random Forest and Support Vector Machines. Other researches have focused on identifying linguistic phenomena related to

✉ Rafael Valencia-García
valencia@um.es

José Antonio García-Díaz
joseantonio.garcia8@um.es

¹ Facultad de Informática, Universidad de Murcia, Campus de Espinardo 30100, Murcia, Spain

¹ [https://www.urbandictionary.com/define.php?term=Poe's Law](https://www.urbandictionary.com/define.php?term=Poe's+Law).

satire, such as sarcasm or irony and recently new resources have been published such as IroSVA 2019 [18].

Regarding the techniques involved on satire identification, the authors of [2] compared traditional machine-learning with deep-learning classifiers in order to outperform the results proposed by [21]. They discover, however, that modern Spanish pre-trained word embeddings do not outperform the results achieved by classical term-counting features. This fact suggests that the presence of certain keywords influences too much in the classification, which may be indicative of certain biases in the corpus. Accordingly, we examined the existing datasets regarding satire identification and discover some biases, most of them related to the distant supervision techniques employed and the limited number of different accounts. In order to address these issues, our contribution is two fold. On the one hand, we present the Spanish SatiCorpus 2021, a balanced dataset for identifying satire with 18,207 satiric utterances and, on the other, we evaluate this dataset with several deep-learning architectures and feature sets, including linguistic features and contextual and non-contextual word and sentence embeddings.

The rest of the manuscript is organised as follows: First, Sect. 2 summarises research focused on satire identification but including also research concerning automatic sarcasm and irony detection, due to their close relationship with satire. Next, Sect. 3 presents and describes the compilation process of the Spanish SatiCorpus 2021. In Sect. 4, the reader can find details as regard of the pipeline used for evaluating the corpus, including the feature sets and the deep-learning architectures involved. The results achieved by each model, and its consequently analysis, are described, respectively, in Sects. 5 and 6. Finally, the conclusions of this research and some promising research directions are listed in Sect. 7.

Background information

First, it is important to remark the similitude and differences between verbal irony, sarcasm, and satire because these terms are sometimes misinterpreted. Satire and sarcasm are both forms of expression. However, whereas satire has moralising purpose in which vices and abuses are caricatured, sarcasm is intended to be hurtful and scathing and it is used as a rhetorical device to subdue an adversary. Both, satire and sarcasm, make use of figurative language to their end. Examples of figurative language are understatements, similes, metaphors, personifications, or idioms among others. Of these literary devices, the verbal irony stands out. Verbal irony consists in that the writer intends to be understood with expressions that contrasts with the literal meaning of what he says. The reader can find a detail overview regarding figurative language, its discriminant features, and techniques employed for its identification in [20].

Regarding sarcasm identification, in [17] the authors present a framework for sarcasm identification based on bidirectional recurrent neural networks and term-weighted trigrams, that they refer as inverse gravity moment. The objective for the term weight strategy is to boost critical words and joint words at the same time that word order is preserved. During their experimentation, the authors compare their framework with models based on word embeddings as well as supervised and unsupervised weighting functions, such as term frequency, odds ratio, balanced distributional concentration or regularised entropy, just to name a few. This framework was evaluated with three sarcasm corpus. The first one is based on Twitter, annotated using distant supervision. The second one is based on the Internet Argument Corpus, and the third one is created from extracting news headlines from satirical news sites such as The Onion.²

In [12], the authors proposed a BERT-based model called ViLBERT and evaluate it against a dataset composed of images and headlines from satirical web sites. This study was conducted from a multi-modal perspective as it incorporates the analysis of images in satirical news that contains images. They observed that these images were created using photo manipulation techniques to create fictional and unreal scenarios. Their results suggests that multi-modal perspective outperforms text-based solutions.

The identification of satire, verbal irony, and sarcasm have been explored in other languages apart from English. Besides, other researches have focused on multi-lingual approaches. The work presented in [9], for instance, focuses on French, English and Arabic for irony detection. It is worth mentioning that the identification of complex language phenomena, such as irony, is heavily cultural dependant. Therefore, the authors compare multicultural languages Indo-European languages with less culturally close languages. Their proposal indicate that monolingual-based models trained from multilingual word representation are beneficial for irony detection in situations in which annotated datasets are not available.

It is possible to find novel research and datasets concerning satire, sarcasm, and irony in other languages apart from English, such as Turkish [16], Bangla [1], or Persian [10]. As far as our knowledge goes, the principal research focused on satire identification in Spanish was performed first by Barbieri [3] and next by Salas-Zárate [21], in which both authors evaluated satire identification from datasets based on satirical and non-satirical headlines. The approach followed by [21] was based on linguistic features extracted from Linguistic Inquiry Word Count (LIWC) [25] whereas the approach followed by [3] compared statistical features and manually crafted linguistic features. It is worth noting that the dataset from [21] were revisited by [2], in which the

² <https://www.theonion.com/>.

authors evaluated other features based on count-term features with traditional machine learning classifiers and word and sentence embeddings from novel Spanish pretrained word embeddings. However, their results show an increment in the accuracy over the Mexican Spanish but not in the European Spanish dataset by employing term-counting features. The non-contextual features achieved lower accuracy in the European Spanish dataset compared with approaches based on linguistic features or term-counting features. The reason of this was that the Mexican Spanish dataset was slightly biased because, on the one hand, there were fewer different Twitter accounts so classifiers can focus on specific linguistic patterns from the community managers and, on the other, as the editorial line of each media is focused on particular topics, these topics biased automatic classifiers. The problems identified for the compilation of dataset related to satire identification are taken into account for the compilation of the Spanish SatiCorpus 2021, which is described below.

Corpus

We rely on the UMUCorpusClassifier tool [6] for extracting satirical and non-satirical texts from Twitter. We named this novel dataset as the Spanish SatiCorpus 2021, a balanced dataset consisted into satirical and non-satirical tweets mostly compiled from news sites (from Europe and Latin America). We select four satirical news sites from Spain, namely, El Mundo Today,³ El Jueves,⁴ El Intermedio,⁵ and Revista Mogolia.⁶ Three from Mexico, namely, El Deforma,⁷ El Dizque,⁸ and El Univerfail.⁹ Finally, there is one more from Venezuela: and the Venezuelan medium El Chigüire Bipolar.¹⁰ In addition, to prevent bias from satirical news sites, we added some accounts from Twitter used for impersonate and satirise real political actors and relevant people.

We use distant supervision for corpus annotation, following a similar criteria as described in [21] and [3], assuming that all documents written for satiric news media are satiric. However, we observe that there are several tweets from the satirical accounts used for promoting events. In order to avoid the problems previously identified in other satirical datasets, we conduct a sanity checking in order to discard tweets that

Table 1 Corpus distribution per label and split

Label	Train	Development	Test	Total
satire	10923	3642	3642	18207
non-satire	10923	3642	3642	18207
total	21846	7284	7284	36414

started from some words such as *entrevista*¹¹ or words related to merchandising. We also remove some clues of the texts, as we found that some media uses hooks such as “INCREIBLE PERO CIERTO”.¹² We also remove those tweets with less than 15 words

One of the problems identified in the previous datasets is that satirical and non-satirical news may not focus on the same events, thus introducing some bias when automatic classifiers can discern among satiric and non-satiric utterances just looking at what the news are about. To mitigate this inconvenient we select the non-satirical tweets in base of the satirical ones. For this, we look for the most similar documents for each satiric document using Text Similarity based on TF-IDF and calculating their cosine distance. We build a matrix in which the rows are the indexes of the satiric documents and the columns the indexes of the non-satiric documents and each component of the matrix represents their distance. Then, we apply an iterative process to select the most similar satiric and non-satiric documents. Once selected, we remove the row and column from the matrix and repeat this process until there are no rows left, that is, we find a match for each satirical document.

The first release of the Spanish SatiCorpus 2021 contains 18,207 satiric and 18,207 non-satiric tweets. This dataset contains tweets between March, 2018 to June, 2021. We divided the dataset into three splits, namely, train, development, and testing, in a ratio of 60-20-20. The dataset is available for the research community at <https://pln.inf.um.es/corpora/satire/spanish-saticorpus-2021.zip>. However, according to the Twitter guidelines,¹³ we only share the identifiers of the tweets as well as the labels and the split (train, development, test) to preserve users' rights over their content. Table 1 contains the statistics per corpus regarding label and split.

Some satiric and non satiric examples from the dataset are depicted in Fig. 1. In the first row, we can compare two news from Mexico. The non-satirical one comments that during Ash Wednesday (a catholic holiday), the ash was distributed in individual bags in order to avoid agglomerations due to the COVID pandemic. However, the satirical counter-news

³ <https://www.elmundotoday.com/>.

⁴ <https://www.eljueves.es/>.

⁵ https://twitter.com/El_Intermedio.

⁶ <https://www.revistamongolia.com/>.

⁷ <https://eldeforma.com/>.

⁸ <https://www.eldizque.com/>.

⁹ <http://www.eluniverfail.com/>

¹⁰ <https://www.elchiguirebipolar.net/>

¹¹ In English: interview.

¹² In English: Incredible, but true.

¹³ <https://developer.twitter.com/en/developer-terms/agreement-and-policy>.

warns citizens not to confuse these sachets with similar ones with inhaled drugs. As we can observe, forcing the datasets to contain similar satiric and non satirical tweets prevents machine learning models to overfit regarding certain keywords and topics related to the domain. In the second row, both satirical and non-satirical news are related to an event that was debated in the congress of the deputies in Spain, in relation to the facilitating mental health treatments in Spain. An open microphone caused that the comment “*Go to the doctor*” from a politician was heard throughout the hemisphere. This fact was criticised in the press because it downplays mental illness issues. However, from a satirical point of view, the news looks like this politician has been a hero, and that his comment saved the life of the politician who was exposing at that time, for something that had nothing to do with mental health.

Materials and methods

In order to evaluate the Spanish SatiCorpus 2021, we conducted a deep-learning pipeline to evaluate it with different feature sets and deep-learning architectures and so, evaluate what the strong features for automatic satire identification are. Figure 2 depicts the architecture of our proposal. In a nutshell, it can be described as follows. First, we apply a text pre-processing stage to clean the dataset (see Sect. 4.1). Second, we divided the dataset into training, validation, and testing in a ratio of 60-20-20 (see Sect. 3). For this, we use a stratified split to ensure the ratio of satiric and non-satiric documents remains balanced. Third, we conduct a feature extraction stage to obtain the linguistic features and the embedding based features (see Sect. 4.2). Fourth, we perform an hyperparameter optimisation stage to evaluate several machine learning models that combines each feature set separately and combined (see Sect. 4.3) and, finally, the best models for each feature set are evaluated with the test dataset.

Text pre-processing

In this stage we remove some parts of the texts that could bias the deep-learning classifiers. First, we remove social media jargon such as hyperlinks, hashtags, or mentions. Besides, we replace digits with a token. Next, to work with features that are based on lexicons, we remove expressive lengthening, that it is a linguistic device used to emphasis some words and we fix misspellings with the PSPELL tool.

It is important to remark that we maintain different versions of the documents to extract some of the linguistic features. For example, to extract features related to Part-of-Speech features and named entities we keep a normalised version but in which the case was preserved.

Feature extraction

In this section we describe the feature sets evaluated. First, we evaluate linguistic features and different kind of embeddings, including pretrained non-contextual word and sentence embeddings from different models as well as pretrained contextual sentence and word embeddings from different versions of BERT.

The **linguistic features (LF)** are extracted using the UMUTextStats tool [7,8]. This tool is inspired in LIWC [25], but designed to the Spanish in mind. It captures a total of 365 different linguistic features organised in the following categories:

- **(COR) Correction and style:** This category checks the correct usage of writing communication. In particular, it distinguishes among (1) orthographic errors, related to the number of misspelled words or the bad use of Spanish accentuation; (2) stylistic errors, related to the presence of sentences that starts with numbers or with the same word; and (3) performance, looking for duplicated words, the usage of dot after exclamation or question signs or the presence of common errors and redundant expressions.
- **(PHO) Phonetics:** This category is related to expressive lengthening, which consists in the elongation of certain letters with an emphasising purpose. As Spanish language contains words in which the same letter appears contiguous twice such as *coordinar*¹⁴ we consider only the repetition of three or more letters and, in the case of vowels, whether or not they have an accent.
- **(MOR) Morphosyntax:** This category includes features that represent how words and sentences are composed. Specifically, there are features that capture (1) grammatical gender, discerning among masculine, feminine, and neutral words; (2) number, to discern among plural and singular; (3) affixes, to capture a fine-grained variety of suffixes (nominals, adjectivizers, verbalizers, adverbializers, augmentative, diminutives, or despective) and prefixes. There are features also organised by PoS categories, such as nouns, verbs, adjectives, adverbs, determiners, pronouns, prepositions, conjunctions, or interjections among others. To capture these features we apply a mixed approach based on Stanza [22], to capture the main categories, and lexicons that capture fine grained subcategories for capturing sub-types.
- **(SEM) Semantics.** This category captures four semantic features: (1) onomatopoeia, to identify words that are composed as regards of the sound they produce. For example, the word “*achís*” in Spanish makes reference to the sound that it is produced when you sneeze; (2) euphemism, that are softer versions of expressions that

¹⁴ In English: to coordinate.

Fig. 1 Spanish SatiCorpus 2021 examples

could be considered too rude in some contexts; (3) dysphemism, that are vulgar words used for replacing other considered more neutral; and (4) synecdoche, that are a type of literary trope used for representing a part as a whole. For example, in “Se quedó con cuatro bocas que alimentar”,¹⁵ the statement “cuatro bocas” refers to four children.

- **(PRA) Pragmatics.** This category captures the presence of figurative language devices, distinguish among understatements, hyperboles, idiomatic expressions, rhetorical questions, verbal irony, metaphors or similes among others. It also contains several linguistic features to see how difference sentences are connected by means of discourse markers. In addition, there are features for capturing some typical Spanish courtesy forms.
- **(STY) Stylometry:** This category captures a (1) wide variety of punctuation symbols, (2) corpus statistics, such as the Token-Type Ratios (TTR), as well as (3) other metrics related to the number of words, syllables or sentences.
- **(LEC) Lexical:** The aim of this category is to capture the topics in the text. For this, we analyse from abstract topics (analytical thinking, achievement, friendship, religion, or certainty among others) to general topics, such as locations, organisations, animals, clothes, food, and a list of professions.
- **(PLI) Psycho-linguistic processes:** This category is related to lexicons and emojis related to sentiments (positive, negative) and emotions (anger, sadness, anxiety).
- **(REG) Register:** This category captures how people use language to communicate, as it could be the presence of informal or cultured language. We also capture topics related to offensive speech.
- **(SOC) Social media jargon:** This category captures features concerning clues that reveals the speaker domains

the social media jargon as it can be specific terminology used on social media or the usage of mechanisms such as hyperlinks, mentions or emojis.

With respect the embeddings, we evaluate non-contextual and contextual word and sentence embeddings. Embeddings are dense vectors that represent words within a latent space. These embeddings are usually learned from unsupervised generic tasks, such as next-word prediction. Word embeddings cluster together words that are semantically similar at the same time they maintain the distance with other clusters of words. Their main drawback, however, is that traditional (non-contextual) embeddings are not aware about polysemy, so the same word have the same representation regardless their significance in a sentence. Contextual words embeddings solve this drawback generating the embeddings by taking into account the context of a word, that is, the words that are next to it. Contextual word embeddings and models based on transformers have meant a great qualitative leap in many NLP tasks, although they are computationally more demanding. This fact is partially mitigated with the usage of sentence embeddings, in which instead of words, sentences are encoded within the latent space. There are different strategies to obtain sentence embeddings, generally taking averaging from the word embeddings. In addition, another of the key-advantages of embeddings, regardless if they are contextual or not, is that they can be learn from generic datasets, which provides two major benefits. One the one hand, pre-trained embeddings already conveys general meaning they converge faster and, on the other, they are a form of transfer learning, in which the embeddings have been learnt with concepts that could not be in the domain.

The **non-contextual word embeddings (WE)** evaluated in this work are based on pretrained models based on word2vec [14], fastText [15], and gloVe [19]. It is worth noting that word embeddings allow to explore specific types of neural networks architectures, such as convolutional and

¹⁵ She was left alone with four mouths to feed.

Fig. 2 System architecture

recurrent neural networks, that are capable for taking profit as regards of the space and temporal dimension of language. That is, convolutional neural networks can generate high-order features by clustering joint words whose significance differs from the one it can be obtained individually, as it can happen with the words *New* and *York*. Recurrent neural networks, on the other hand, explore the temporal dimension, so they take into account the order of the words. Particularly, we evaluate two bidirectional recurrent neural networks based on Long-Short Term Memory (LSTM) and Gated Recurrent Units (GRU). The **non-contextual sentence embeddings (SE)** are extracted from FastText [11], which Spanish model is trained from Wikipedia and CommonCrawl.

The **contextual word embeddings (BETO)** are evaluated from the Spanish version of BERT [4]. These contextual embeddings are focused on Spanish. In addition, we have compared their reliability with **multi-lingual embeddings from BERT (mBERT)** [5] and its **distilled version (dmBERT)** [24]. For all these embeddings we rely on the HuggingFace transformers library to fine tune the model with the Spanish SatiCorpus 2021. However, these kind of embeddings are difficult to combine with other feature sets as they are time consuming. Therefore, for the **contextual sentence embeddings (BF)** we extracted the fixed representation of the [CLS] token as suggested in [23] and we use this representation to combine the contextual embeddings more easily

with the rest of the feature sets. Our results suggest that the performance of contextual sentence and word embeddings are similar in terms of precision, recall, and accuracy, with this and other datasets.

After the feature sets were extracted, we conduct a feature selection and a feature normalisation stage. First, we apply a MinMax scaler to the linguistic features as they contains features in different scales with raw counts and percentages. Next, we select the best features using Information Gain, discarding those features that belong to the last quartile.

Hyper-parameter tuning

An hyper-optimisation stage is conducted in order to evaluate the configuration of several neural networks architectures for satire identification. Our strategy consist into train a independent neural network per feature set. We rank the models by accuracy. The hyperparameters evaluated are: (1) the number of hidden layers as well as the number of neurons per layer; (2) the dropout rate, to avoid overfitting; (3) the activation function; (4) the learning rate and the usage of a time-based decay scheduler; and (5) the batch size.

The best architecture and hyperparameters for feature set can be seen in Table 2. It can be appreciated that the best accuracy is always obtained with shallow neural networks, that is, neural network with one of two hidden layers maximum. We observe this fact both in individual feature sets (LF, SE, WE, BF), or when combined in groups of two, three or four within the same neural network. However, the number of neurons per layer vary achieving better results with larger number of neurons in models in which the LF are present. We notice that the number of neurons with WE is small, with only 2 neurons. A similar finding results when combining SE and BF, with 4 neurons. Concerning the dropout rate, we observe than only four experiments provided better results without it: WE, BF with SE, and LF, SE, WE, and BF. With respect to the activation function, ReLu is the one that appears more often, regardless when the feature sets are in isolation or combined.

Results

For each model we evaluate the precision (see Eq. 1), recall (see Eq. 2), and $F1$ -score (see Eq. 3) of the satirical and non-satirical class. In addition, we include the macro-averaged scores of both labels together with the accuracy (see Eq. 4) to compare the overall performance of each model.

$$\text{Precision} = TP / (TP + FP) \quad (1)$$

$$\text{Recall} = TP / (TP + FN) \quad (2)$$

$$F1_Measure = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

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

Table 3 depicts the results for each feature set. It can be observed that the best result is achieved with BF (96.78% of accuracy), which accuracy outperforms largely the rest of the feature sets: LF (85.14% of accuracy), WE (84.25% of accuracy), and SE (82.04% of accuracy). We can observe that LF are good indicators for satire identification, improving the results from SE and WE. The precision between satiric and non-satiric sentences is similar with all feature sets, although we can observe in SE the highest difference, observing larger precision for satiric documents. A similar but slighter behaviour is appreciated in case of multi-lingual BERT. When comparing the results achieved with different version of transformers (BETO, multi-lingual BERT and distilled multi-lingual BERT), learning the embeddings from an specific language rather than a multilingual corpus is beneficial, as there is difference of accuracy of 6.37% between BETO and mBERT, and a difference of accuracy of 13.042% between BETO and distilled mBERT. This finding is inline with other research that compare embeddings learnt from monolingual or multilingual data sources [13,26]. The accuracy achieved with BF is similar to the achieved with the pretrained model from BETO (96.897% of accuracy). Thus, the fixed vectors from the [CLS] token extracted for the fine-tuned model contains the relevant information of the sentence without sacrificing performance. Moreover, we can observe from Table 2, that the best results were achieved with a shallow neural network composed of one hidden layer of 48 neurons. However, this fixed sentence representation eases their combination with other feature sets improving the performance.

Next, Table 4 shows the results achieved by combining different feature sets in the same neural network, applying a knowledge integration strategy. LF improves the results when combined with SE (89.044% of accuracy), WE (87.740% of accuracy) and BF (97.405% of accuracy). This finding indicates that the linguistic features contains additional information that the one than can be extracted from the embeddings. Other combinations based only in embeddings, however, do not improve their results significantly, as it can be observed from SE and WE (84.308%) and SE with BF (96.870%). It is worth mentioning that accuracy achieved with the combination of LF, SE, and BF (97.32% of accuracy) is lower to the one obtained from combining LF and BF. The combination of all features (LF, SE, WE, and BF) is also slower (97.268%) but it achieves very good precision and recall with both satiric and non-satiric documents.

Next, we evaluate the combination of the models by means of ensembles. These results are shown in Table 5. Specifically, three ensemble learning strategies were evaluated: (1) highest probability, (2) weighted mode, and (3) the average of probabilities. The highest probability consist into inspect

Table 2 Hyperparameter evaluation over the validation dataset per feature set in isolation or combined within the same neural network

Features	Architecture	Shape	\# of layers	\# of neurons	Dropout	lr	Activation
lf	Dense	Brick	2	48	0.2	$1e-2$	Sigmoid
se	Dense	Brick	2	256	0.1	$1e-3$	ReLU
we	Dense	Brick	1	2	None	$1e-3$	ReLU
bf	Dense	Brick	1	48	0.3	$1e-3$	ReLU
lf-se	dense	Brick	1	512	0.3	$1e-2$	ReLU
lf-we	Dense	Brick	1	128	0.2	$1e-2$	ReLU
lf-bf	Dense	Brick	2	128	0.3	$1e-2$	Tanh
se-we	Dense	Brick	1	64	0.2	$1e-3$	None
se-bf	Dense	Brick	1	4	None	$1e-2$	None
we-bf	Dense	Brick	1	64	None	$1e-3$	ReLU
lf-se-bf	Dense	Brick	2	48	0.2	$1e-2$	Tanh
lf-se-we	Dense	Brick	2	256	False	$1e-3$	ReLU
lf-we-bf	Dense	Brick	2	16	0.3	$1e-2$	ReLU
lf-se-we-bf	Dense	Brick	2	256	False	$1e-3$	ReLU

the probability reported by each model that one text belongs to the satirical or non satirical category and select the highest one. As we can observe, this strategy reports the highest precision for the non-satire label and the highest recall for the satire class. Therefore, this strategy can be suited in systems that focus on one of these metrics. However, both the recall of the non-satire class and the precision of the satire class are limited. Next, the weighted mode strategy consists in that each model emits a vote for determining whether a text is satiric or not. This strategy is also known as soft voting. As not all classifiers have the same, we ranked them according to their results with the validation split, so the vote for BF is slightly superior to the rest. Note that the weighted mode strategy achieves the overall best result compared with the rest of strategies evaluated, reporting a macro weighted *F1*-score of 95.510%, performing well regardless of the metric evaluated (precision or recall) and regardless the label (satiric or non satiric). The last strategy evaluated consisted in to average the predictions of all classifiers and so emit the final vote. This strategy achieves slightly lower results than the weighted mode, with a macro weighted *F1*-score of 93.712125% but with similar performance in both labels and metrics.

Analysis

The Spanish SatiCorpus 2021 has been evaluated applying different feature sets in isolation (see Table 3) or combined by means of ensembles or within the same neural network (see Tables 4 and 5, respectively). These results are all very competitive. We observe that there is a significant quality leap with contextual word embeddings, regardless if they are sentence or word-based. Moreover, contextual embed-

dings based on BETO [4] improves significantly the results achieved by embeddings learnt from multi-lingual corpus (mBERT). We also observe that the LF in isolation achieved superior accuracy than the models based on non-contextual embeddings, improving SE (85.145% of accuracy with LF vs 82.042% of accuracy with SE) and WE (85.145% of accuracy with LF vs 84.253% of accuracy with WE). In addition, a relevant finding is that the combination of the LF with other kind of ensembles results in more reliable systems. However, the combination of non-contextual ensembles does not outperform the results achieved separately. We observe a significant increment of the combination of the LF with SE (89.044% of accuracy) and with WE (87.740% of accuracy). However, the increment with BF was less significant (from 96.787% to 97.405% of accuracy).

It is worth mentioning that the training of neural networks adopting the knowledge integration strategy is a time-eating task. First of all, because we perform a extra hyperparameters optimisation process for these neural networks instead of combining the best architectures individually. Secondly, because the neural networks that rely on word embeddings (as opposite of sentence embeddings) requires millions of parameters rather than the thousands evaluated with sentence embeddings. In contrast, the ensemble learning strategy is faster, as we combine the predictions of already existing models, at the same time their results are only slightly lower (from an accuracy of 97.323% combining LF, SE, and BF to 95.511%, with an ensemble of LF, SE, WE, and BF applying the weighted mode).

As the linguistic features provide some kind of interpretability, we calculate the contribution of each linguistic category separately. This information is shown in Table 6. We can observe that stylometry (STY) is the linguistic category that provides major reliability regarding precision and recall

Table 3 Precision, recall, *F1*-score of satiric and non-satiric labels by feature set separately

Feature set	Precision	Recall	<i>F1</i> -score	Accuracy
(LF) Linguistic features				
Non-satire	86.240	83.635	84.918	–
Satire	84.115	86.656	85.367	–
Macro-avg	85.178	85.146	85.142	85.146
(SE) Non-contextual sentence embeddings				
Non-satire	79.831	85.750	82.685	–
Satire	84.609	78.336	81.352	–
Macro-avg	82.220	82.043	82.018	82.043
(WE) Non-contextual word embeddings				
Non-satire	83.983	84.651	84.316	–
Satire	84.528	83.855	84.190	–
Macro-avg	84.255	84.253	84.253	84.253
(BETO) Contextual words embeddings				
Non-satire	97.550	96.211	96.876	–
Satire	96.262	97.584	96.918	–
Macro-avg	96.906	96.897	96.897	96.897
(mBERT) Contextual words embeddings				
Non-satire	88.398	93.300	90.783	–
Satire	92.907	87.754	90.257	–
Macro-avg	90.652	90.527	90.520	90.527
(dmBERT) Contextual words embeddings				
Non-satire	85.411	81.658	83.492	–
Satire	82.430	86.052	84.202	–
Macro-avg	83.920	83.855	83.847	83.855
(BF) Contextual sentence embeddings				
Non-satire	97.228	96.321	96.772	–
Satire	96.355	97.254	96.802	–
Macro avg	96.792	96.787	96.787	96.787

Macro-averaged precision, recall, *F1*-score, and accuracy of the overall result

(76.3130% of precision, 76.126% of recall). We observe that the feature categories that contains fewer linguistic features, that are, phonetics (PHO) and semantics (SEM), are the ones that achieved lower results, with an 34.115% and a 34.366% of *F1*-score respectively. In fact, when observe the results achieved per class, it can be noticed that the recall of these models were biased to the non satirical category predicting this feature almost a 100% of the times. The reason of this is that the LF does not appear very often in the dataset. For example, only a small proportion of the documents make use of elongation (PHO), so any model is not capable of making reliable predictions with only this information.

As we can observe, the results achieved with the LF categories merged outperforms largely the results achieved separately, achieving an accuracy of 85.142% (see Table 3). This fact indicates that the linguistic categories are comple-

Table 4 Precision, recall, *F1*-score of satiric and non-satiric labels by feature sets combined

Label	Precision	Recall	<i>F1</i> -score	Accuracy
(LF + SE)				
Non-satire	89.217	88.825	89.020	–
Satire	88.874	89.264	89.068	–
Macro-avg	89.045	89.044	89.0444	89.044
(LF + WE)				
Non-satire	89.531	85.475	87.456	–
Satire	86.105	90.005	88.012	–
Macro-avg	87.818	87.740	87.734	87.740
(LF + BF)				
Non-satire	97.945	96.842	97.391	–
Satire	96.878	97.968	97.420	–
Macro-avg	97.411	97.405	97.405	97.405
(SE + WE)				
Non-satire	84.545	83.965	84.254	–
Satire	84.074	84.651	84.362	–
Macro-avg	84.310	84.308	84.308	84.308
(SE + BF)				
Non-satire	97.549	96.156	96.847	–
Satire	96.210	97.584	96.892	–
Macro-avg	96.879	96.870	96.870	96.870
(WE + BF)				
Non-satire	97.347	96.733	97.039	–
Satire	96.753	97.364	97.058	–
Macro-avg	97.050	97.048	97.048	97.048
(LF + SE + WE)				
Non-satire	88.061	88.303	88.182	–
Satire	88.271	88.029	88.150	–
Macro-avg	88.166	88.166	88.166	88.166
(LF + WE + BF)				
Non-satire	97.108	96.815	96.961	–
Satire	96.825	97.117	96.971	–
Macro-avg	96.966	96.966	96.966	96.966
(LF + SE + BF)				
Non-satire	97.782	96.842	97.310	–
Satire	96.872	97.803	97.336	–
Macro-avg	97.327	97.323	97.323	97.323
(LF + SE + WE + BF)				
Non-satire	97.281	97.254	97.268	–
Satire	97.255	97.282	97.268	–
Macro-avg	97.268	97.268	97.268	97.268

Macro-averaged precision, recall, *F1*-score, and accuracy of the overall result

mentary. For more clarity, we calculate the average of all features organised by linguistic category and label (satire and non-satire) in order to observe this differences in a polar chart (see Fig. 3). It draws our attention than the average

Table 5 Precision, recall, *F1*-score of satiric and non-satiric labels by ensemble learning strategy for LF, SE, WE, and BF

Label	Precision	Recall	<i>F1</i> -score	Accuracy
Highest probability				
Non-satire	99.752	66.310	79.664	–
Satire	74.769	99.835	85.503	–
Macro-avg	87.260	83.072	82.583	83.072
Weighted mode				
Non-satire	95.623	95.387	95.505	–
Satire	95.399	95.634	95.516	–
Macro-avg	95.511	95.511	95.511	95.511
Probability average				
Non-satire	93.332	94.152	93.740	–
Satire	94.100	93.273	93.685	–
Macro-avg	93.716	93.712	93.712	93.712

Macro-averaged precision, recall, *F1*-score, and accuracy of the overall result

of each linguistic category is similar regardless the class, with the notable exception of correction and style (COR), in which non-satiric documents are most relevant. As it can be observed, there are no data regarding semantics (SEM), phonetics (PHO), and pragmatics (PRA). Regarding psycholinguistic process (PLI) we observe no significant difference and only small difference concerning stylometry (STY). With respect to the usage of social media jargon (SOC) and register (REG), there are differences among satiric and non-satiric utterances although these categories do not appear regularly in the dataset.

Next, we calculate the contribution of each linguistic feature to the class by obtaining the mutual information. We select the top most discriminatory features and we observed if they appear more often in satirical or non-satirical documents. This information is depicted in Fig. 4. We can observe than features related to readability (STY), that involves the number of words, syllables and sentences are strong discriminatory features. However, in general, these features have a similar representation for satiric and non-satiric documents. Surprisingly, from correction and style category (COR), we observe than the number of orthographic errors are more common in non satirical documents than in satirical documents. A similar finding is observed with the number of hashtags, that appear mostly in non-satirical documents. Concerning morphological features, the use of pronouns and nouns are discriminatory features, being slightly more frequently the pronouns in satirical documents whereas nouns are more common on non-satirical documents. This finding suggests that satirical documents makes use of personification to perform the satire.

Table 6 Ablation analysis per linguistic category

Linguistic category	Precision	Recall	<i>F1</i> -score	Accuracy
(COR) Correction and style				
Non-satire	63.686	57.688	60.539	–
Satire	61.330	67.106	64.088	–
Macro-avg	62.508	62.397	62.313	62.397
(PHO) Phonetics				
Non-satire	50.172	99.973	66.813	–
Satire	96.296	0.714	1.417	–
Macro-avg	73.234	50.343	34.115	50.343
(MOR) Morphosyntax				
Non-satire	73.539	60.132	66.163	–
Satire	66.280	78.364	71.817	–
Macro-avg	69.909	69.248	68.990	69.248
(SEM) Semantics				
Non-satire	50.062	99.121	66.525	–
Satire	56.164	1.126	2.207	–
Macro-avg	53.113	50.124	34.366	50.124
(PRA) Pragmatics				
Non-satire	51.490	87.781	64.907	–
Satire	58.605	17.298	26.712	–
Macro-avg	55.047	52.540	45.810	52.540
(STI) Stylometry				
Non-satire	74.095	80.340	77.091	–
Satire	78.531	71.911	75.075	–
Macro-avg	76.313	76.083	76.484	76.126
(LEC) Lexical				
Non-satire	62.529	58.649	60.527	–
Satire	61.065	64.854	62.903	–
Macro-avg	61.797	61.752	61.715	67.752
(PLI) Psycho linguistics processes				
Non-satire	54.430	53.130	53.772	–
Satire	54.224	54.519	54.864	–
Macro-avg	54.327	54.325	54.318	54.325
(REG) Register				
Non-satire	54.366	82.235	65.457	–
Satire	63.549	30.972	41.647	–
Macro-avg	58.957	56.604	53.552	56.604
(SOC) Social media jargon				
Non-satire	67.512	50.439	57.740	–
Satire	60.443	75.728	67.227	–
Macro-avg	63.977	63.083	62.484	63.083

Precision, recall, *F1*-score of satiric and non-satiric labels and the macro-averaged precision, recall, *F1*-score, and accuracy of the overall result

Comparison with previous datasets

In order to compare the reliability of our proposal, we compare our methods with the dataset compiled by Barbieri [3],

Fig. 3 Polar chart per linguistic category arranged by satiric and non satiric documents

the two datasets compiled by Salas-Zarate [21] concerning European Spanish and Mexican Spanish, and the three datasets for the IroSVA 2019 shared task [18] based on European Spanish and Spanish from Cuba and Mexico. Table 7 contains the results for each related feature set. We compare the results with our approach based on the combination of LF with BF, applying a knowledge integration strategy, as this approach provide our best results with the Spanish SatiCorpus 2021. It is important to remark that neither [21] nor [3] provided the distribution of the labels for training and testing as they rely on ten-cross fold validation. However, in order to maintain our pipeline, we divided these datasets into training, validation, and testing similar as we do with the Spanish SatiCorpus 2021.

The first experiments conducted by Barbieri [3], achieved a macro $F1$ -score of 85.200% by using linguistic features and hand-craft features. Our approach achieves a macro $F1$ -score of 90.007%. However, these results should be compared with caution, as Barbieri report their results with ten-cross validation whereas we report us with a test set of 20%. We observed that the most relevant linguistic features are related to stylometry and punctuation, including the readability and the length of the documents. We also observe that common nouns and suffixes were relevant morphosyntax features.

When compared with Salas-Zarate [21], we can observe than for the European Spanish dataset our approach based on LF and BF largely outperforms the results achieved with LIWC, from an 85.5% of accuracy to a 95.6477% of accu-

racy. In this sense, the results achieved by Salas-Zarate using only linguistic features are competitive, but far for the performance that contextual embeddings provides. In case of the Mexican Spanish dataset, our results also outperform the ones from Salas-Zarate, from a 84% of accuracy to a 92.8427%. In case of the European Spanish, we found that morphological features stand out from the stylometric clues for the classification, including indicative simple present and imperative verbs tenses and prepositions, but also the presence of orthographic mistakes. In case of Mexican Spanish, however, due to the biases of the dataset, the social media features (hashtags, mentions) are the most relevant features. There are also relevant features related to misspellings and stylometry and punctuation.

Finally, we compare our approach with IroSVA 2019. Note that as this competition is recent, the approaches submitted by the participants of the task were very competitive, and our approach only outperforms the results achieved with the Cuban dataset with a macro $F1$ -score of 66.3360%, compared with the macro $F1$ -score of 65.9600%, which was the official best result of the task. In case of Spanish and Mexican, however, the results were slightly inferior. 70.9440% vs 71.6700% of macro $F1$ -score with European Spanish, and 66.3950% vs 68.03% of macro $F1$ -score with Mexican Spanish. We found some differences among the linguistic features among all datasets. First, we only found as relevant performance errors for the Spanish, whereas we could not identify any kind of correction and style relevant feature for Mexican and Cuban. It also draws our attention the presence of psycho linguistic process in Mexican and Cuban. However, these features were positive for the Mexican and related to anger in Cuban. Besides, we found a small bias in the Spanish dataset with the presence of hyperlinks as relevant feature. The most common linguistic category among all datasets is stylometry, with features related to the length of the documents (Spanish), the TTR standard Token (Mexican), and the number of sentences (Cuban).

After the analysis of the results and the comparison with other datasets, we achieve the following insights:

- The results achieved by contextual sentence and word embeddings are similar. However, sentence embeddings are easier and faster to combine with other features.
- The usage of embeddings learnt from monolingual approaches are superior from the ones learnt with multilingual datasets. as the results achieved with the Spanish version of BERT outperform largely the results achieved with multilingual BERT.
- The identification of satire by means of linguistic features are more confident with the usage of stylometric clues rather than other clues such as semantics, pragmatics, or by sentiment.

Fig. 4 Information gain of the top ranked linguistic features per label

Table 7 Comparison of accuracy and macro averaged precision, recall, and $F1$ -score with other datasets related to satire and irony identification

Dataset	Approach	Spanish variant	M-P	M-R	M- $F1$ -score	Accuracy
Satire [3] (2015)	SVM	European	–	–	85.200	–
	LF + BF	European	90.049	90.037	90.001	90.037
Satire [21] (2017)	LIWC	European	85.500	85.500	85.500	85.500
	LF + BF	European	95.602	95.655	95.627	95.648
	LIWC	Mexican	84.600	84.000	84.000	84.000
	LF + BF	Mexican	92.845	92.862	92.842	92.843
Irony [18] (2019)	–	European	–	–	71.670	–
	LF + BF	European	70.530	72.375	70.944	72.833
	–	Mexican	–	–	68.030	–
	LF + BF	Mexican	67.292	65.897	66.395	71.333
	–	Cuban	–	–	65.960	–
	LF + BF	Cuban	67.555	65.750	66.336	71.500

- Linguistic features are complementary with all kind of embeddings, regardless if they are word or sentence-based, and regardless if they are contextual or non-contextual. However, the contribution of the linguistic features with the contextual sentence embeddings is limited, due to the high performance achieved with these embeddings in isolation.
- There is no one linguistic category that stands out from the rest. Although the best results for each characteristic were obtained with the stylometry category, the general results obtained by LF are superior. We also observed that some categories are not relevant for satire identification, specially those related to phonetics (PHO), semantics (SEM), and pragmatics (PRA).

Conclusions and further work

In this paper we have described the compilation of the 2021 SatiCorpus 2021 dataset for satire identification. This corpus is balanced and contains a total of 36414 documents. During its compilation, we focused on avoid the problems identified when we analysed past resources that were related to the limitation of the number of different accounts, the heterogeneity between the topics discussed on satirical and non-satirical documents, and the presence of clues within the texts that can biased the results in order to identify the author of the tweet rather than if the document is satirical or not. As commented in Section 3, this dataset has been released for the scientific community. In addition, this dataset has been com-

prehensively analysed with different feature sets and their combinations, including linguistic features and a wide variety of embeddings.

As future promising research lines, we suggest to extend the Spanish SatiCorpus 2021 to include information regarding what mechanisms for performing satire are involved. On the one hand, this proposal is harder to resolve than simply binary classification. On the other hand, these insights will provide major understanding concerning satire identification. We also propose to investigate the interpretability of the linguistic features avoiding model-agnostic approaches, and focusing of how they perform within the neural network, in order to gain better understanding about in which cases LF and embeddings are complementary. Regarding the selection of the model, we will evaluate to change the hyperparameter optimisation stage, based on random search, for other approaches such as Bayes optimisation. In this sense, we will evaluate other strategies to compare the models and the adoption of nested-cross validation strategies to obtain models that generalise better.

Acknowledgements This paper is part of the research project LaTe4PSP (PID2019-107652RB-I00) funded by MCIN/AEI/10.13039/501100011033. In addition, José Antonio García-Díaz is supported by Banco Santander and the University of Murcia through the Doctorado industrial programme.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Al Imran A, Wahid Z, Ahmed T (2020) Bnnet: a deep neural network for the identification of satire and fake Bangla news. International conference on computational data and social networks. Springer, Berlin, pp 464–475
- Apolinario-Arzube Ó, García-Díaz JA, Medina-Moreira J, Luna-Aveiga H, Valencia-García R (2020) Comparing deep-learning architectures and traditional machine-learning approaches for satire identification in Spanish tweets. *Mathematics* 8(11):2075
- Barbieri F, Ronzano F, Saggion H (2015) Is this tweet satirical? A computational approach for satire detection in Spanish. *Procesamiento del Leng Nat* 55:135–142
- Cañete J, Chaperon G, Fuentes R, Ho JH, Kang H, Pérez J (2020) Spanish pre-trained bert model and evaluation data. In: PML4DC at ICLR 2020
- Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 4171–4186
- García-Díaz JA, Almela Á, Alcaraz-Mármol G, Valencia-García R (2020) Umucorpusclassifier: compilation and evaluation of linguistic corpus for natural language processing tasks. *Procesamiento del Leng Nat* 65:139–142
- García-Díaz JA, Cánovas-García M, Palacios RC, Valencia-García R (2021) Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings. *Future Gener Comput Syst* 114:506–518. <https://doi.org/10.1016/j.future.2020.08.032>
- García-Díaz JA, Cánovas-García M, Valencia-García R (2020) Ontology-driven aspect-based sentiment analysis classification: an infodemiological case study regarding infectious diseases in Latin America. *Future Gener Comput Syst* 112:641–657. <https://doi.org/10.1016/j.future.2020.06.019>
- Ghanem B, Karoui J, Benamara F, Rosso P, Moriceau V (2020) Irony detection in a multilingual context. *Adv Inf Retr* 12036:141
- Golazizian P, Sabeti B, Asli SAA, Majdabadi Z, Momenzadeh O, Fahmi R (2020) Irony detection in persian language: a transfer learning approach using emoji prediction. In: Proceedings of The 12th Language Resources and Evaluation Conference, pp 2839–2845
- Krasnowska-Kieraś K, Wróblewska A (2019) Empirical linguistic study of sentence embeddings. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp 5729–5739
- Liu, R., & Zhou, X. (2021, August). Grenzlinie at SemEval-2021 Task 7: Detecting and Rating Humor and Offense. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021) pp. 281–285
- Libovický J, Rosa R, Fraser A (2019) How language-neutral is multilingual bert? arXiv preprint [arXiv:1911.03310](https://arxiv.org/abs/1911.03310)
- Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint [arXiv:1301.3781](https://arxiv.org/abs/1301.3781)
- Mikolov T, Grave E, Bojanowski P, Puhresch C, Joulin A (2018) Advances in pre-training distributed word representations. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)
- Onan A, Toçoğlu MA (2020) Satire identification in Turkish news articles based on ensemble of classifiers. *Turk J Electr Eng Comput Sci* 28(2):1086–1106
- Onan A, Toçoğlu MA (2021) A term weighted neural language model and stacked bidirectional lstm based framework for sarcasm identification. *IEEE Access* 9:7701–7722
- Ortega-Bueno R, Rangel F, Hernández Farias D, Rosso P, Montes-y Gómez M, Medina Pagola JE (2019) Overview of the task on irony detection in Spanish variants. *Proc Iber Lang Eval Forum* 2421:229–256
- Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 1532–1543
- del Pilar Salas-Zárate M, Alor-Hernández G, Sánchez-Cervantes JL, Paredes-Valverde MA, García-Alcaraz JL, Valencia-García R (2020) Review of English literature on figurative language applied

- to social networks. *Knowl Inf Syst* 62(6):2105–2137. <https://doi.org/10.1007/s10115-019-01425-3>
21. del Pilar Salas-Zárate M, Paredes-Valverde MA, Rodríguez-García MÁ, Valencia-García R, Alor-Hernández G (2017) Automatic detection of satire in twitter: a psycholinguistic-based approach. *Knowl Based Syst* 128:20–33. <https://doi.org/10.1016/j.knosys.2017.04.009>
 22. Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD (2020) Stanza: a python natural language processing toolkit for many human languages. In: Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations, pp 101–108
 23. Reimers N, Gurevych I, Reimers N, Gurevych I, Thakur N, Reimers N, Gurevych I (2019) Sentence-BERT: sentence embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on empirical methods in natural language processing. Association for computational linguistics
 24. Sanh V, Debut L, Chaumond J, Wolf T (2019) Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv* [arXiv:abs/1910.01108](https://arxiv.org/abs/1910.01108)
 25. Tausczik YR, Pennebaker JW (2010) The psychological meaning of words: Liwc and computerized text analysis methods. *J Lang Soc Psychol* 29(1):24–54
 26. Virtanen A, Kanerva J, Ilo R, Luoma J, Luotolahti J, Salakoski T, Ginter F, Pyysalo S (2019) Multilingual is not enough: Bert for finnish. *arXiv preprint* [arXiv:1912.07076](https://arxiv.org/abs/1912.07076)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.