Social Network Analysis and Mining (2023) 13:43
https://doi.org/10.1007/s13278-023-01049-0

ORIGINAL ARTICLE

Exploiting stacked embeddings with LSTM for multilingual humor and irony detection

Radiathun Tasnia[1] · Nabila Ayman[1] · Afrin Sultana[1] · Abu Nowshed Chy[1] · Masaki Aono[2]

Received: 21 August 2022 / Revised: 15 February 2023 / Accepted: 17 February 2023 / Published online: 3 March 2023

Abstract

Humor and irony are types of communication that evoke laughter or contain hidden sarcasm. The opportunity to diversely express people's opinions on social media using humorous content increased its popularity. Due to subjective aspects, humor may vary to gender, profession, generation, and classes of people. Detecting and analyzing humorous and ironic emplacement of informal user-generated content is crucial for various NLP and opinion mining tasks due to its perplexing characteristics. However, due to the idiosyncratic characteristics of informal texts, it is a challenging task to generate an effective representation of texts to understand the inherent contexts properly. In this paper, we propose a neural network architecture that couples a stacked embeddings strategy on top of the LSTM layer for the effective representation of textual context and determine the humorous and ironic orientation of texts in an efficient manner. Here, we perform the stacking of various fine-tuned word embeddings and transformer models including GloVe, ELMo, BERT, and Flair's contextual embeddings to extract the diversified contextual features of texts. Later, we use the LSTM network on top of it to generate the unified document vector (UDV). Finally, the UDV is passed to the multiple feed-forward linear architectures for attaining the final prediction labels. We present the performance analysis of our proposed approach on benchmark datasets of English and Spanish language. Experimental outcomes exhibited the preponderance of our model over most of the state-of-the-art methods.

Keywords Humor · Irony · Stacked embeddings · Flair · BERT · Feed-forward linear architecture

1 Introduction

Nowadays, the emerging trends of using online social media platforms bring the exponential growth of user-generated content, where users are prompt to ubiquitous utilization of figurative and creative language, like humor and irony. Humor, as well as irony, is widespread linguistic phenomena. Frequently, humor is interconnected with irony though they have different origins. Through social media platforms (e.g., Facebook and Twitter), people share their humorous or ironic opinions or emotions on controversial or crucial issues with emojis, short texts, symbols, and images. The automatic distinction of humor and irony in texts is one of the most important and beneficial tasks due to its various significant applications including social media analysis, sentiment analysis, opinion mining, product reviews, and human-centered artificial intelligence (AI) domain (Pannu 2015).

Humor is a complex and multifaceted phenomenon that has been studied from a variety of perspectives, including psychological, sociological, and philosophical perspectives. From the psychological viewpoint, humor can be defined

Radiathun Tasnia, Nabila Ayman, Afrin Sultana have contributed equally to this work.

✉ Radiathun Tasnia
radia.tasnia.cu@gmail.com

Nabila Ayman
nabila.ayman.cu@gmail.com

Afrin Sultana
afrin.sultana.cu@gmail.com

Abu Nowshed Chy
nowshed@cu.ac.bd

Masaki Aono
aono@tut.jp

[1] Department of Computer Science and Engineering, University of Chittagong, Hathazari, Chittagong 4331, Bangladesh

[2] Department of Computer Science and Engineering, Toyohashi University of Technology, 1-1 Hibarigaoka, Tempaku-cho, Toyohashi, Aichi 441-8580, Japan

as the cognitive and emotional reaction to stimuli that are perceived as amusing or comical (Meyer 2000). According to sociological research, the most commonly examined aspect of humor appreciation is how it aligns with cultural norms and practices (Reyes et al. 2012a). Humor from a cognitive perspective refers to the psychological and neural processes involved in the perception and comprehension of humorous stimuli (Brône et al. 2006; Brône 2017). From a pragmatic perspective, humor refers to the study of how it is used and understood in human communication and interaction (Hoicka 2014).

Therefore, humor is a form of communication that evokes laughter, smiles, or various emotions, and can be used to express ideas, feelings, and thoughts. Emphasis on multiple word senses, cultural knowledge, subjectivity, and pragmatic competence of humorous text poses interesting linguistic challenges to natural language processing (NLP). Besides, the perception of a humorous text can differ from person to person due to age, gender, and socioeconomic status. As a consequence, humor controversy arises when a humorous text portrays offensive contexts to a specific group of people. Automatic detection of humorous and controversial humorous text from a pile of data is indeed an arduous task. The first conference on computational humor was organized in 1996 (Hulstijn 1996); then, numerous works (Reyes et al. 2012b; van den Beukel and Aroyo 2018; Weller and Seppi 2019) have been conducted on humorous text identification tasks. Some subtasks bring novelty by combining humor detection with offense (Meaney et al. 2021) or identifying the influence of figurative languages in sentiment analysis (Ghosh et al. 2015). Most recently, Meaney et al. (2021) have introduced a shared task in SemEval-2021 to tackle humor and the subjectivity of humor appreciation. This task consists of two subtasks and each one comprises a number of tasks. Here, the first subtask is a humor detection task which is composed of a binary task to predict whether a given text is humorous or not, a regression task to predict the level of a humorous text, and another binary task to predict the humor controversy based on the classified humor. On the contrary, the second subtask is intended to predict the level of offensive text in the range of 0 to 5. In this paper, we evaluate the performance of our system for the humor and humor controversy identification tasks.

Besides English, several works also have been conducted in Spanish on the research area of humor recognition. Castro et al. (2018) introduced a task at IberEval-2018 to detect humor along with its funniness score prediction. The second version of this task (Chiruzzo et al. 2019) was organized in the following year at IberLEF. It proposed the same task formats as the previous year, a binary task to predict whether tweets are a joke or not, and determines the level of funniness if the tweet contains a joke in it. Most recently, Chiruzzo et al. (2021) have shared a task at IberLEF-2021 providing a large dataset of humor in Spanish for the researchers to gain a better insight into the direction of analyzing humor structure. This task is partitioned into four subtasks for analyzing the humorous characteristics that exist in the tweets. A binary task to predict Spanish tweets as humorous or not, rating the humorous tweets, finding the humor mechanism from a set of classes, and predicting the content of the humor based on its targets. In this paper, we also focus on the Spanish humor detection task.

Irony is a form of speech or writing where the intended meaning is different from the literal meaning (Sperber and Wilson 1981; Wilson and Sperber 1992; BREDIN 1997). Rather than this, we can define irony from various perspectives. Irony from a psycholinguistic perspective refers to a phenomenon where a speaker's words or actions have a meaning that is different or opposite to their intended meaning. This difference can be recognized by listeners who infer the ironic meaning through various cues such as tone of voice, context, and situational factors (Jr et al. 1995). The comprehension of irony is believed to heavily rely on context and is thought to involve complex inferential processes beyond literal interpretation (Colston and Gibbs Jr 1998; Giora and Fein 1999). Irony from a pragmatic perspective can be defined as a form of language use that involves saying one thing while meaning the opposite, to achieve a particular communicative goal, such as criticism or humor (CLARK 1984; Wilson 2006; Wilson and Sperber 2012). From a cognitive viewpoint, irony is the result of intricate inferential processes that arise from conflicting conceptual scenarios (Ruiz de Mendoza and Lozano 2021; Lozano-Palacio and de Mendoza Ibáñez 2022). The cognitive viewpoint recognizes the role of the mental processes of audiences in understanding irony. Hence, irony is a dynamic and interactive phenomenon that involves both the speaker and the listener (Tobin and Israel 2012; Peña and Ruiz de Mendoza 2017).

Moreover, irony is the act of echoing an idea that has been attributed to a person, a group, or the public at large while also expressing mockery or hostility toward the idea (Wilson 2006). The ironist displays an attitude toward this proposition and others who may hold or have held it by employing ironic language rather than the actual meaning of the proposition or its opposite. In line with pretense theory, when a speaker makes an ironic statement, they are engaging in a form of "mental play" in which they and the listener temporarily adopt an alternate perspective to understand the irony (Clark and Gerrig 1984). From a psychological standpoint, the listener's comprehension of an ironic remark depends heavily on the shared ground that the ironist and the audience have, such as their shared ideas, knowledge, and presumptions.

Classifying a piece of text as ironic or non-ironic is a challenging task due to its rhetorical trope and contextual incongruity. Prior research (Reyes et al. 2013; Buschmeier

Table 1 Examples of independence relationship of humor and irony

Text	Class label
E#1: I never finish anything. I have a black belt in partial arts	Humorous and ironic
E#2: I can't believe I forgot to go to the gym today. That's 7 years in a row now	Humorous and non-ironic
E#3: I love waking up at 8 am on a Saturday morning after going to bed at midnight	Non-humorous and ironic
E#4: Learn from the scars of others	Non-humorous and non-ironic

et al. 2014) has been performed on irony detection to scrutinize its twisted nature. Later, Van Hee et al. (2018) introduced a shared task in SemEval-2018 to address the challenges of determining ironic tweets. This task is combined with two subtasks where subtask A is intended to detect the ironic nature of tweets and subtask B is deliberated to determine the categories of ironic tweets (verbal irony using a polarity contrast, other verbal ironies, situational irony, and non-irony). In this paper, we evaluate our proposed method to detect ironic tweets.

Irony and humor fall under distinct levels of classification. We may encounter both humorous and non-humorous ironies and both types of humor as they are independent phenomena. To understand the independent behavior of these phenomena, we have listed some examples in Table 1. The first example (E#1) contains a humorous and ironic context. It is humorous because it makes fun of the idea of someone having a black belt in "partial arts," a play on the phrase "martial arts" and ironic because the speaker says they have a "black belt" in "partial arts," implying that they have achieved mastery in not finishing things. This is contrasting with the typical expectation of a black belt as a symbol of mastery or achievement in a specific skill or discipline. The second example (E#2) expresses funniness because it creates a paradoxical situation that is unexpected and has a comic effect. The humor comes from the unexpected and exaggerated length of time, as well as the lighthearted way the speaker presents this information. The sentence is not ironic because it does not involve a discrepancy between what is said and what is meant. Instead, it is a straightforward statement with a humorous twist. The third example (E#3) is ironic but not humorous because it is a contrast between expectation and reality. On a Saturday morning, one would expect to have a leisurely sleep-in, and the idea of waking up early is not something that is normally associated with this day. The speaker's statement of "loving" waking up early on a Saturday morning is a contrast between what one would expect and what is being expressed, and this contrast creates an ironic effect. The sentence is not humorous, as humor typically involves a playful or comical aspect that is absent in this sentence. The last example (E#4) is a literal sentence that is neither humorous nor ironic.

Numerous studies have been done on multimodal irony (Tomás et al. 2022) and multimodal humor (Brône 2021; Hasan et al. 2021) in addition to verbal irony and humor. Multimodal humor and multimodal irony refer to the use of multiple media of communication, such as verbal, visual, acoustic, and gestural to convey a humorous or ironic message. This combination of different modes can make the message more impactful as it allows the speaker to convey multiple layers of meaning and to play with the audience's expectations. The multimodal functionalities allow users to communicate their humorous thoughts by combining material in different ways and diverse formats. However, in this study, we employed textual humor and irony to conduct our work.

Table 2 articulates examples of humorous, non-humorous, and controversial sentences from SemEval-2021 HaHackathon, ironic and non-ironic expressions from SemEval-2018 irony detection, and Spanish humorous and non-humorous sentences from IberLEF-2021 HAHA shared tasks. Here, the first example (E#1) contains only funny context, so it is a humorous sentence that contains no controversy about its humor. The second example (E#2) itself expresses funniness, but the second part of the example produces contention making its funniness unpleasant to a specific group of people. Therefore, the second example (E#2) is a humorous sentence with controversy. In the fourth example (E#4), the positive word "loving life" conveys a twisted meaning in the whole sentence. Therefore, it is an ironic sentence. The fifth example (E#5) is not ironic as the words of the example are well suited to their essence and have no twisted interpretation. We also include humorous and non-humorous examples from the Spanish language in E#6, and E#7, respectively. For the analysis purpose, we utilize Google Translate to translate the Spanish samples to their equivalent English text. The sixth (E#6) example implies the meaning as "Mosquitoes that practice Formula 1 in your ear." which comprises funny content to depict the pragmatic scenario of day to days life. So, it falls under the humorous category. On the contrary, the seventh (E#7) example delineates the meaning as "When the cashier checks my bills to see that they are not fake, I do the same with the ones she gives me in change, so she can see what it feels like." that represents a contradictory phenomenon from a logical viewpoint rather than expressing humor. As a consequence, this sentence categorizes as non-humorous.

To tackle the challenges of above mentioned humor and irony detection tasks, researchers proposed different kinds of

Table 2 Examples of humorous, controversial humorous, and ironic texts

Text	Class label
Examples from SemEval-2021 HaHackathon task	
E#1: Damn girl! Your name must be Ebola... All I can think about is you spreading	Humorous and non-controversial
E#2: I never finish anything. I have a black belt in partial arts	Humorous and controversial
E#3: Learn from the scars of others	Non-humorous and non-controversial
Examples of irony from SemEval-2018 irony detection task	
E#4: 3 h sleep yay loving life	Ironic
E#5: My whole life is just "oh ok"	Non-ironic
Examples of humor in Spanish from IberLEF-2021 HAHA task	
E#6: Mosquitos que practican Formula 1 en tu oreja	Humorous
E#7: Cuando la cajera revisa mis billetes para ver que no son falsos, hago lo mismo con los que me da de cambio, para que vea lo que se siente	Non-humorous

approaches. Numerous studies (Reyes et al. 2012b; van den Beukel and Aroyo 2018; Tasneem et al. 2020) address the task challenges through employing conventional machine learning classifiers using a wide range of handcrafted features including lexical, semantic, and syntactic features. However, extracting various hand-engineered features is a tedious job, hence it is a hassle itself and also not feasible (González-Ibánez et al. 2011). Deep learning methodologies can capture hidden dependencies between terms. The combination of pre-trained word embeddings (e.g., GloVe, ELMo) with deep learning neural architectures (e.g., CNN, LSTM) was introduced and managed to gain remarkable performance on different humor and irony identification tasks (Amir et al. 2016). Later, various studies focused on transformer-based architectures due to their effectiveness over traditional deep learning approaches (Tay et al. 2020). But utilizing a single transformer does not effectively explore the diversity of contextual features. To tackle these aforementioned drawbacks, we proposed a stacked embedding-based approach to explore the diversity of contexts of a text.

The key contribution of this paper is that we exploit a fine-tuned stacked embeddings approach to combine various word and transformer embeddings to capture diverse word semantics in context. Utilizing LSTM architecture on word and transformer embeddings, we construct a unified document vector (UDV) that efficiently demonstrates the syntactic and semantic properties of a document to derive global context. To acquire better insights into an intermediate representation, we feed the unified document vector in multiple feed-forward linear architectures and procure our final predictions from the last layer. The utilization of lightweight features from the last layer gives better delineation of text and makes our system more robust and memory efficient. We also furnish a precise comparative performance analysis with state-of-the-art approaches in English and Spanish based on five benchmark datasets. Experimental results demonstrate the efficacy and generalization of our approach.

The rest of this paper is structured as follows: Sect. 2 states related work including both conventional and deep learning methods on humor and irony detection. Section 3 presents the detail of our proposed framework. In Sect. 4, we illustrate our experimental setup and discuss the performance evaluation of our approach. Section 5 inspects the fallacy, robustness, and explainability of our system. With an outlook on future research direction, we conclude our work in Sect. 6.

2 Related work

The colossal appearance of user-generated humor, offensive, and ironic texts on online platforms drive the necessity of identifying these contents and increase its demands in the domain of emotional intelligence. All these identification processes are relying on propensity, sentiments, and emotions. However, detecting humor, irony, and offensive content turns into a challenging task as humor and irony create complex structures where offense gradually contaminates the social platforms. In recent days, researchers are giving more focus to the automatic detection of these indicators for solving various NLP and linguistic challenges (Van Hee et al. 2018; Chiruzzo et al. 2019; Zampieri et al. 2019; Meaney et al. 2021).

Humor and offense are deliberated on a large scale from the perspective of intellectual, emotional, and semantic viewpoints exploring their hidden structures. Nowadays, researchers are more interested in experimenting from a computational perspective, which not only gives better perceptions of their definition but also extends the area of humor and offensive fabrication and their ingredients. Humor detection describes the process of diagnosing farcical and jocular contents from the text wherein offensive detection indicates whether the content is derogatory and obnoxious. Several works have been accomplished for detecting

humorous and offensive language. Reyes et al. (2012b) proposed various sets of features by exploiting polarity and emotional content with classifiers for detecting humor. van den Beukel and Aroyo (2018) conducted homophones and homographs as features with classifier settings, whereas Khandelwal et al. (2018) annotated the language and humorous tags orchestrating various features and classifiers. They consolidated N-gram, common words, and hashtags in the features branch, whereas they integrated kernel SVM, random forest, extra tree, and naïve Bayes in the classifiers segment (Khandelwal et al. 2018). Weller and Seppi (2019) explored transformer-based neural network architecture using the ratings from Reddit pages to detect humor labels. Davidson et al. (2017) used a crowd-sourced hate speech lexicon, labeling tweets with a multiclass classifier for differentiating between hate speech and offensive text.

At the 2019 13th International Workshop on Semantic Evaluation (SemEval-2019), Zampieri et al. (2019) introduced a task OffensEval that was aimed to distinguish between offensive and non-offensive Twitter posts. In this task, Liu et al. (2019a) used the fine-tuned bidirectional encoder representation from transformers (BERT). Expanding in the multimodal field, at the SemEval-2020, Sharma et al. (2020) ventilated a task memotion analysis that focused on sentiment analysis, overall emotion classification, and classifying emotion intensity of memes, whereas Swamy et al. (2019) composed the ensemble of six different classifiers for detecting offensive language. Swamy et al. (2020); Sharma et al. (2020) followed the approach including a logistic regression baseline, a BiLSTM with the attention-based learner, and a transfer learning approach with BERT for detecting humor from the multimodal-based system. Recently, at the SemEval-2021, Meaney et al. (2021) have unveiled a task HaHackathon that pivoted on binary humor detection, prediction of humor and offense ratings, and humor controversy task. In this task, Song et al. (2021) utilized various fine-tuned pre-trained language models including RoBERTa and ALBERT stacking with a simple linear regression model, whereas Faraj and Abdullah (2021) deployed diverse fine-tuned pre-trained models including RoBERTa, BERT, ALBERT, and XLNet with hard-voting ensemble technique for humor and humor controversy detection.

Besides English, diverse works have been conducted for automatic humor detection in Spanish and researchers proposed various methods for analyzing humor in this multilingual field. At the IberLEF-2019, Chiruzzo et al. (2019) introduced the HAHA task which was based on automatic detection and rating of humor in Spanish tweets. At this task, Ismailov (2019) combined BERT-base multilingual cased with fastai library,[1] binarized multinomial naïve Bayes with unigram–bigram TF-IDF features and logistic regression, whereas Mao and Liu (2019) only utilized BERT-base multilingual cased where they counted the last layer output including the [CLS] token and linear output. Altin et al. (2019) focused on a multitask supervised learning module unifying humor, sentiment, irony, and aggressiveness where they merged dialect-specific word embeddings, a common BiLSTM layer, and two dense layers as classifiers. Giudice (2019) followed a different approach in the training phase. They amalgamated a character label 1-D CNN with three layers integrating a Bi-RNN and a dense layer. Miller et al. (2019) employed Gaussian process preference learning with various formatted handcrafted features. They fused three-word renditions which are the average token frequency in a Wikipedia dump, Spanish Twitter embeddings, and the word's lemma average polysemy. Cattle et al. (2019) exploited a document tensor space for embedding tweets, trained with random trees classifier where they contemplated each tweet as a document.

Along with this direction, several works have also been done in the domain of irony detection. Irony detection refers to the most intense genre as it categorizes the bizarreness depending on the medium context. It inclined to various polarities for conveying the particular scenario. Previous tasks on irony detection are largely based on statistical machine learning and neural network methods prioritizing several rules and features. Rule-based methods identify ironic tweets capturing lexical and sequence-based context. Some of them choose hashtags as their sequence taggers for instance (Liebrecht et al. 2013; Reyes et al. 2013). They triggered "#sarcasm" and "#irony" hashtags for detecting sarcasm and irony, respectively. Maynard and Greenwood (2014) implemented hashtags for identifying sarcasm and utilized hashtags result for sentiment purposes. Recently, researchers are mostly focused on deep learning-based approaches to innately capture the ironic contexts. Various neural architecture-based approaches are observed deploying different kinds of models. Amir et al. (2016) applied CNN, whereas Zhao et al. (2018) employed the composition of CNN with BiLSTM. In this direction, if we talk about features-based statistical machine learning methods for irony detection, a vast portion of work is regulated through handcrafted features. Among them, Barbieri and Saggion (2014) conducted their experiment on lexicon-based features, whereas Joshi et al. (2015) applied lexical features. Some others address this challenge by combining various features and classifiers-based modules (Tasneem et al. 2020). Huang et al. (2018) considered false positive hashtags for irony identification, whereas Bharti et al. (2015) conducted

[1] https://docs.fast.ai/

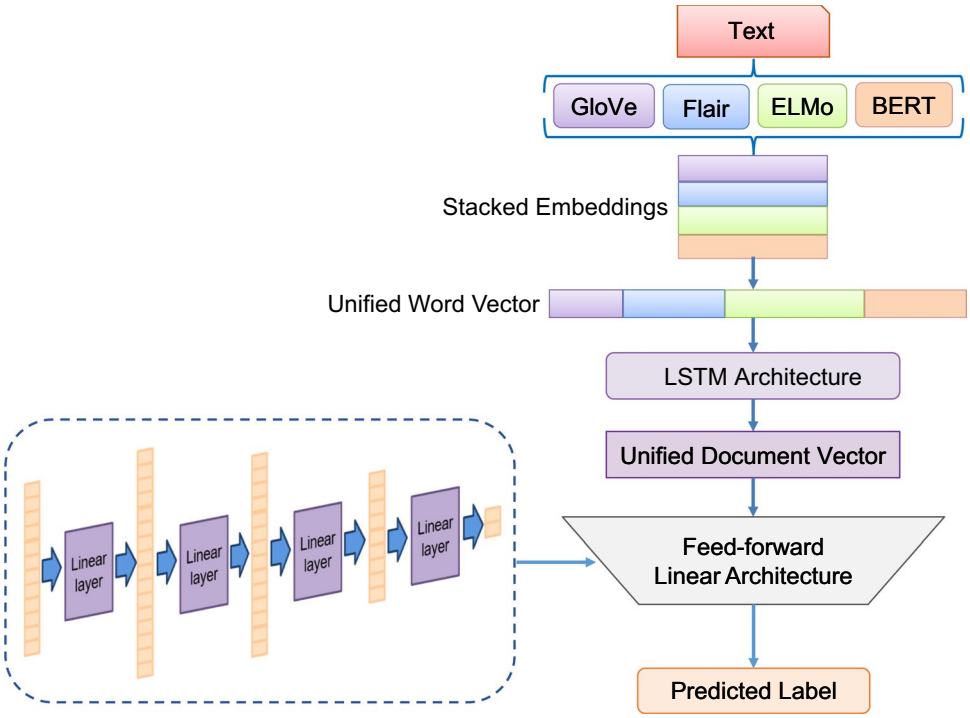


Fig. 1 Proposed framework

double approaches for sarcasm detection: One is negative sentiment-based and another is adverb and adjective captured exclamation.

Later, to tackle the challenges of irony detection, Van Hee et al. (2018) organized a task at the SemEval-2018 which resolute the existence and type of ironic tweets. In this competition, González et al. (2018); Wu et al. (2018) utilized a combination of LSTM with CNN and densely connected LSTM, respectively. Besides, Pamungkas and Patti (2018) focused on rhetorical features, whereas Rohanian et al. (2018), Pamungkas and Patti (2018) considered some other features including syntactical and sentiment-aware features. Augmenting the features pathway, Joshi et al. (2015) focused on implicit congruity, pragmatic, and explicit congruity features. Based on the extracted features, researchers exploited a various number of classifiers for training purposes. Most of the works regulated on SVM classifiers (Barbieri and Saggion 2014; Rohanian et al. 2018; Pamungkas and Patti 2018) and many researchers also implemented logistic regression and XGBoost classifiers (Rangwani et al. 2018; Rohanian et al. 2018).

In brief, we have perceived that most of the researchers explored the traditional approaches in their proposed methods including an ensemble of various preprocessing techniques, handcrafted features, and statistical classifiers. Most recently, transformer-based models obtain popularity among researchers and people applying various transformer-based models including BERT, RoBERTa, ALBERT, etc. Transformer-based models understand the sentence context effectually. However, employing a single transformer-based model may not capture the diverse contextual information from a textual segment which motivates us to address an effective method for detecting humor, humor controversy, and irony. To overcome the limitation of extracting diverse contexts of words, we employ stacked embeddings of the various fine-tuned transformer and word embedding models. We merge these embeddings features by utilizing document LSTM embeddings to generate a unified document vector which later feeds in multiple feed-forward linear architectural frames and exploits the last feature level for attaining the final prediction labels. Moreover, diversification in embeddings feature, as well as segmentation of feature vectors, obtains good results not only in the SemEval-2021 humor, humor controversy detection, and SemEval-2018 irony detection tasks but also in IberEval-2018, IberLEF-2019, and IberLEF-2021 Spanish humor detection tasks following multilingual settings.

3 Proposed framework

Our proposed framework aims to determine humor and irony labels from the informal user-generated text. The overview of our proposed framework is depicted in Fig. 1.

For a given input text, we extract embedding features using various models including GloVe, ELMo in word embeddings, BERT in transformer word embeddings, and Flair embeddings. To generate effective word representation, we combine the extracted embedding features through the stacked capsule and feed them to an LSTM architecture to

obtain a unified document vector. The merged feature vectors apply to the multiple feed-forward linear architectures splitting them into several feature levels. We employ the last feature zone to procure the final prediction labels. In the rest of the descriptions of our manuscript, we refer to this method as StackedEmbedding_LA.

3.1 Word embeddings

Word embeddings is a set of language modeling and feature generation techniques where individual words are represented as real-valued vectors in a pre-defined vector space. A vector of a real number with many dimensions represents each word. In our model, we use GloVe, ELMo, contextual Flair embeddings, and BERT word embeddings. We briefly discuss the aspect of these embedding models in our proposed architecture in the following subsections.

3.1.1 GloVe

Global Vectors for Word Representation (GloVe) (Pennington et al. 2014) is an unaided learning algorithm that captures words in a vectorized form. In the training phase, vector space occupies a linear substructure that accumulates global word–word co-occurrence statistics from the corpus. Employing GloVe, we can capture both local and global contextual information of words. It deploys the matrix factorization technique where we utilize word context as a matrix. We employ GloVe word embeddings in the stacked module of our architecture for extracting both local and global contextual representations of text to understand the ironic and humourous context effectively.

3.1.2 Contextual flair embedddings

Flair's library (Akbik et al. 2018) of contextual word embeddings helps us to embed words depending on their contextual use and pre-trained on large unlabeled corpora. Here, contextual use means the complex characteristic of words for which a particular word can have different meanings depending on the context. For example, we can see that the word "break" has different meanings in two different situations:

*Why did you **break** my watch?*

*We should take a coffee **break** for 20 minutes.*

When we process humor from the text, we can notice the contextual use of words. We know that human language can be very divertive. Flair embeddings help in finding the contextual meaning of a particular word in a sentence. It can produce embeddings based on the polysemous use of the word. For these characteristics, Flair embeddings can improve the performance of a model in detecting humor, humor controversy, and irony from given texts.

Contextual Flair embeddings composed of forward and backward models. The forward model tries to predict the next word of the sequence, whereas the backward model tries to predict the preceding word of the sequence. Flair news embeddings trained with one billion English word corpus.[2] Flair also offers Flair es embeddings that is trained with Wikipedia for the Spanish Language.[3] We stack Flair embeddings for English and Spanish texts to cover better perception of the context in a sequence.

3.1.3 ELMo

ELMo stands for Embeddings from language models which is developed by AllenNLP (Peters et al. 2018). ELMo achieved state-of-the-art results in various NLP tasks (Ilić et al. 2018; Peters et al. 2018). ELMo embeddings are deep contextualized and character-based representations of the word. We combine ELMo in stacked embeddings so that we can achieve the deep contextual representation of words which in turn helps to improve the performance of humor and irony detection tasks.

It works with bidirectional language models (biLM). This biLM has two layers. These layers work with a forward pass and a backward pass and are pre-trained over a large text corpus. That is how the layers learn about the context. As a result, this language model ultimately gives an intermediate word vector for each word of the sentence. ELMo is different from the other word embeddings in the way that the traditional word vectors represent a word with a random vector and sometimes fail to capture the word context properly. On the other hand, ELMo captures the deep contextual representation of a word and also takes care of its subjective use. There are different pre-trained models for ELMo. We use the pre-trained model which is in the English language (Peters et al. 2018).

3.1.4 Transformer word embeddings

While the stacked embeddings system is giving us the chance to combine various embedding models, we tried to use the best models to get the best combination. In recent times, BERT (Devlin et al. 2018) has achieved outstrip performance in many NLP tasks. The main benefit of this model is that it applies the bidirectional training of the transformer to language modeling. This characteristic allows the model

[2] https://github.com/flairNLP/flair/blob/master/resources/docs/embeddings.

[3] https://github.com/flairNLP/flair/blob/master/resources/docs/embeddings.

to learn the context of a particular word based on all of its surroundings and thus can predict the next word accurately. BERT-base model uses 12 layers of transformer encoders. The output of each token from each layer of these encoders works as word embeddings. Traditional word embeddings as Word2Vec produce identical representations for each word, even so, the context within which the word appears can be different. Contrastingly, BERT produces such vector representations of words that are dynamically affected by the words around them. To attain transformer embeddings of each word, we employ the BERT-base uncased (Devlin et al. 2018) model for the English texts and BETO (Canete et al. 2020), a Spanish BERT for the Spanish texts. The size of BETO is similar to the BERT-base model. It is trained on a big Spanish corpus of three billion words with the Whole Word Masking technique.

3.2 Stacked embeddings

Stacked embeddings is a special kind of approach that combines multiple embeddings to build a powerful word representation model without much complexity. It allows the mix and match of embeddings. Stacked embeddings incorporate multi-input models to form the final word vectors to represent a word. It produces a single word vector with a concatenation of different input models to capture the benefits of contextual diversity. The representation of the stacked embedding-based framework is depicted in Fig. 1. We unify GloVe, Flair, ELMo, and BERT embeddings utilizing the stacked embeddings approach of the FLAIR framework to concatenate different embeddings of the same lexical word. It captures both the semantic and syntactic information of the word. The stacked representation of these embeddings can be defined as follows:

$$\mathbf{w}_i = \begin{bmatrix} \mathbf{w}_i^{\text{GloVe}} \\ \mathbf{w}_i^{\text{Flair}} \\ \mathbf{w}_i^{\text{ELMo}} \\ \mathbf{w}_i^{\text{BERT}} \end{bmatrix}$$

where $\mathbf{w}_i^{\text{GloVe}}$, $\mathbf{w}_i^{\text{Flair}}$, $\mathbf{w}_i^{\text{ELMo}}$, and $\mathbf{w}_i^{\text{BERT}}$ denote the GloVe, Flair, ELMo, and BERT embeddings feature vectors, respectively.

3.3 Document embeddings

After capturing the stacked embeddings of different word embedding models of an individual word as described in Sect. 3.2, the unified word vectors are fed to an LSTM network to produce a unified document vector (UDV). This UDV is the vector representation for each sentence using the stacked word embeddings of every word of that sentence. We discuss our used LSTM network in the following subsection.

3.3.1 Document LSTM embeddings

LSTM (Hochreiter and Schmidhuber 1997) stands for long short-term memory network. It is a type of recurrent neural network. LSTM generates the estimated hidden vector sequence and the output vector sequence. Generally, an LSTM unit is composed of a cell. The cell recognizes values over arbitrary time intervals. LSTM is composed of a memory cell, input gate (I_g), output gate (O_g), and forget gate (F_g). These gates are used to determine the transformations of the recent memory cell, C_g and the recent hidden state, H_g. At each time step, the output of the module is controlled by these gates as a function of the old hidden state H_{g-1} and the input at the current time step X_g. The LSTM transition functions are defined as follows (Zhou et al. 2015):

$$I_g = \phi(W_j \cdot [H_{g-1}, X_g] + B_j)$$
$$F_g = \phi(W_f \cdot [H_{g-1}, X_g] + B_f)$$
$$Q_g = tanh(W_q \cdot [H_{g-1}, X_g] + B_q)$$
$$O_g = \phi(W_o \cdot [H_{g-1}, X_g] + B_o)$$
$$C_g = F_g \odot C_{g-1} + I_g \odot Q_g$$
$$H_g = O_g \odot tanh(C_g)$$

Here, ϕ is the logistic sigmoid function that has an output in [0, 1], *tanh* denotes the hyperbolic tangent function that has an output in [-1, 1], and $\odot$ denotes the elementwise multiplication. To understand the mechanism behind the architecture, F_g is viewed as the function to control to what extent the information is going to be thrown away from the old memory cell. I_g is the value of new information that is going to be stored in the current memory cell, and O_g is the output based on the memory cell C_g.

Document LSTM embeddings run an LSTM-type RNN over all words of a sentence and use the final state of the networks as embeddings for the whole document. We stack different kinds of word embeddings that produce $\mathbf{w}_i$ stacked embeddings. We pass this $\mathbf{w}_i$ embeddings to the document LSTM embeddings to get M-dimensional document embeddings of each sentence where M is the hidden size of LSTM architecture.

3.4 Classification module

An M-dimensional unified document vector (UDV) generated from LSTM architecture is fed into a stacked structure of four linear layers to classify each text. The linear layer module applies a linear transformation to the input using its

stored weights and biases. The equation of each feed-forward linear layer can be defined as follows:

$$Z_i = W_i[W_{i-1} * Z_{i-1} + B_{i-1}] + B_i, \text{ Here } i \in [0, 3]$$
$$\text{if } i = 0, \text{ then } W_{i-1} = 1,\ Z_{i-1} = x \text{ and } B_{i-1} = 0.$$

where x is the M-dimensional document feature vector, Z_i is the output feature vector of the i^{th} linear layer, and W_i and B_i are the input weight and bias of the i^{th} layer, respectively.

The first fully connected layer transforms the M-dimensional document feature vector into a k-dimensional hidden vector that passes to the second linear layer as input to produce a p-dimensional feature vector. The third feed-forward linear layer provides an n-dimensional lightweight feature vector from the p-dimensional vector, which is then passed to a fully connected softmax layer to generate the prediction label. Procuring the final prediction labels through utilizing lightweight features provides a better delineation of the sentence and captures the sentence context more effectually than heavyweight features. The Softmax activation function normalizes the feature vector into probabilities as follows:

$$P(Y_i|Z_i) = \text{softmax}(Z_i) = \frac{e^{Z_i}}{\sum_{i=0}^{1} e^{Z_i}}$$

Here, a class with the highest probability value is considered as the final label for each input text.

4 Experiments and evaluation

4.1 Dataset collection

To assess the performance of our proposed StackedEmbedding_LA method, we utilized several benchmark datasets on different tasks. At first, we employ the dataset released at the SemEval-2021 Task 7 (Meaney et al. 2021) for humor and humor controversy classification. The organizers collected 10,000 texts from Twitter and the Kaggle Short Jokes dataset to conserve diversification. 20 annotators, aged 18–70 annotated the dataset. The given training set consists of 8000 texts where 4932 texts are humorous and these 4932 humorous texts are further annotated as controversial and non-controversial. So, the training set contains 2465 controversial and 2467 non-controversial humorous texts. The development dataset comprises 1000 texts, of which 632 are humorous and 385 are non-humorous. Among 632 humorous sentences, 308 sentences contain controversiality. However, we also utilized the non-humorous sentences for humor controversy classification by treating them as non-controversial texts. The test set also comprises 1000 texts.

Besides, to determine the utility of our system for humor classification in Spanish, we employed the large dataset released at the IberLEF-2021 HAHA (Chiruzzo et al. 2021) task. The dataset consists of 36,000 tweets where 24000 tweets are available for training that is biased to the non-humorous class as it contains 14747 non-humorous tweets. The development set possesses 6000 tweets including 2342 humorous and 3658 non-humorous tweets. The rest 6000 tweets are accessible for testing in which humorous and non-humorous tweets are equally distributed. To distinguish the appropriacy of our system, we also employed the dataset of HAHA task at IberEval-2018 (Castro et al. 2018) and IberLEF-2019 (Chiruzzo et al. 2019) that contain 16000 and 24000 tweets as a train set, and each with 4000 tweets as a test set, respectively. Due to the unavailability of the development set, we utilized 10% of the training set as a development set.

Besides, we made use of the dataset of SemEval-2018 Task 3 (Van Hee et al. 2018) to observe the efficacy of our system on irony detection. The dataset contains 4618 tweets including 3834 tweets for training and 784 tweets for the test. Here, we utilized 10% of the training set as a development set. The statistics of our used datasets are illustrated in Table 3.

Moreover, the irony dataset captured high noisy content, whereas the humor dataset seized low noisy content in text. So, we preprocessed the irony dataset and retained the originality of the humor dataset both in English and Spanish. In our preprocessing modules, we employed six different types of preprocessing techniques. We used the approach reported in Baziotis et al. (2017) for splitting the hashtag into meaningful words to understand its semantic context. We utilized Emot[4] for demonizing emojis into textual form to clarify the different contexts of the text. For converting non-standard words into a standard form, we employed two normalization dictionaries reported in Han et al. (2012), Liu et al. (2012). We removed stop words employing the approach reported in Loper and Bird (2002) as these words have not contributed much to the classification module. For ensuring the similarities between the context of the same words, we stemmed modified words to their traditional form implying Krovetz.[5] Accented characters create a discrepancy in a text context, so we removed these accented characters utilizing Unicode[6] and generalized the character format in the text. Here, we depict the examples of applied preprocessing techniques in Table 4.

To train our StackedEmbedding_LA system for the above mentioned tasks, we used the train set and utilized the development set for hyper-parameter tuning. In the end, we assess our system employing the test set.

[4] https://github.com/NeelShah18/emot.

[5] https://github.com/rmit-ir/KrovetzStemmer.

[6] https://pypi.org/project/unicode/.

Table 3 Statistics of used dataset

Task	Category	Train	Dev	Test	Total
Humor detection	*SemEval-2021 task 7*				
	Humorous	4932	632	615	6179
	Non-humorous	3068	368	385	3821
	Total	8000	1000	1000	10000
Humor controversy detection	*SemEval-2021 task 7*				
	Controversial	2465	308	279	3052
	Non-controversial	2467	324	336	3127
	Total	4932	632	615	6179
Irony detection	*SemEval-2018 task 3*				
	Ironic	1720	191	311	2222
	Non-ironic	1731	192	473	2396
	Total	3451	383	784	4618
Humor detection in Spanish	*IberLEF-2021 HAHA task*				
	Humorous	9253	2342	3000	14595
	Non-humorous	14747	3658	3000	21405
	Total	24000	6000	6000	36000
	IberLEF-2019 HAHA task				
	Humorous	8328	925	1492	10745
	Non-humorous	13273	1474	2508	17255
	Total	21601	2399	4000	28000
	IberEval-2018 HAHA task				
	Humorous	5279	586	1492	7357
	Non-humorous	9122	1013	2508	12643
	Total	14401	1599	4000	20000

Table 4 Applied preprocessing techniques' example

Technique name	Original tweet	Preprocessed tweet
Hashtag Segmentation	Just paid $2.59 for gas! #ThanksObama #sarcasm	just paid $2.59 for gas! #thanksobama #sarcasm thanks obama sarcasm
Stemming	3 episodes left I'm dying over here	3 episode left i'm die over here
Stop Word Removal	Just great when you're mobile bill arrives by text	great mobile bill arrive text
Demojize	What are you doing 😂	what are you doing face with tears of joy
Accented Character Removal	You have to coördinate with your mentor	you have to coordinate with your mentor

4.2 Evaluation metric

We utilized the several benchmark datasets released in the shared tasks at the SemEval, IberEval, and IberLEF workshops as mentioned in Sect. 4.1. Therefore, to evaluate the performance of our method, we used the evaluation metric used in these tasks. At the SemEval-2021 humor and humor controversy detection task, the F1 score was applied as the primary evaluation metric. Moreover, at the IberLEF-2021, IberLEF-2019, and IberEval-2018 Spanish humor detection tasks, the F1 score for the humorous category was considered as the main metric. On the contrary, at the SemEval-2018 irony detection task, the F1 score for the positive class (only) was used as a primary evaluation metric. However, accuracy, precision, and recall were appraised as secondary evaluation metrics in all of the mentioned tasks. It is difficult to judge a model signifying one parameter only because a model may work well in one parameter but poor in others. So, it is necessary to evaluate the model based on multiple parameters. For expanding the validity of our proposed method, we also reported the performance of our proposed model categorizing accuracy, precision, and recall scores along with the F1 score for each task.

Table 5 Model configuration for hyper-parameter settings

Task	Hyper-parameter	Configuration
Humor and humor controversy detection	learning_rate	4e-5
	mini_batch_size	16
	anneal_factor	0.8
	max_epochs	6
Irony detection	learning_rate	3e-5
	mini_batch_size	16
	anneal_factor	0.8
	max_epochs	50
Humor detection in Spanish	learning_rate	3e-5
	mini_batch_size	16
	anneal_factor	0.5
	max_epochs	2

4.3 Model configuration

We used Google Colab's GPU for the training and parameter tuning of our system. We present the configuration of our best performing systems in Tables 5 and 6.

In the embedding section, we deployed various embedding combinations both for word embeddings and document embeddings. We amalgamed BERT, RoBERTa, GloVe, ELMo, XLNet, GPT, Flair news-forward, and Flair news-backward for word embeddings. For document embeddings, we trialed the FLAIR implementation of DocumentLSTMEmbeddings and DocumentRNNEmbeddings. We obtained our best configuration through stacking GloVe, ELMo, Flair news-forward, and BERT with DocumentLSTMEmbeddings. The configuration of our best performing system represents each text into a 5988-dimensional feature vector unifying a 100-dimensional transfer learning feature vector from GloVe, a 2048-dimensional word embedding vector from Flair news-forward, a 3072-dimensional feature vector from ELMo, and a 768-dimensional feature vector from BERT-base uncased as referred in Sect. 3.2. We utilized Spanish Flair and Spanish BERT for automatic humor detection in Spanish where Flair "news-forward" reinstated with Flair "es-forward" and "bert-base-uncased" supplanted with "dccuchile/bert-base-spanish-wwm-uncased" (Cañete et al. 2020). For DocumentLSTMEmbeddings as described in Sect. 3.3, we experimented with hidden size and reproject word dimensions 512, 256, 312, 100, and 50. We fixed hidden size 512 and reproject word dimension 256 which depicts the number of hidden states in LSTM and output dimension of reprojected token embeddings, respectively. We fixed reproject word as true which defines a Boolean value and indicates whether to reproject the token embeddings in a separate linear layer before applying them to the LSTM module.

To select the optimal hyper-parameters in individual word embeddings, we trialed with various parameter values. For BERT, we trialed with the top three, four, five, and bottom three, four, and five layers. We found our best result for -1, -2, -3, and -4 which means the top four layers, and averaged these layers. For ELMo, we experimented with layers "all," "top," "average" and embedding size "small," "medium," "original." We observed our best result in layers "all" indicates concatenation of the three ELMo layers with embedding size "original" defines 4096 hidden sizes, 2 layers with 93.6 M parameters.

However, in the classification module, as described in Sect. 3.4, the $M = 512$-dimensional document feature vector obtained from the LSTM module passed to four feed-forward linear layers. We employed a simple random grid search to select the optimal hidden sizes and empirically set the hidden sizes of the feed-forward layers as $k = 2000$, $p = 1500$, and $n = 50$ dimensions. The last linear layer is the fully connected softmax activation layer. Besides, we perform hyper-parameter tuning on learning rate, anneal factor, mini-batch size, and max epochs. We conducted some experiments on each of these hyper-parameters. We

Table 6 Model configuration for embedding settings

Task	Embeddings	Settings
Humor and irony detection	ELMoEmbeddings	"original," "all"
	TransformerWordEmbeddings	"bert-base-uncased," layer="-1,-2,-3,-4," layer_mean=True
	FlairEmbeddings	"news-forward"
	WordEmbeddings	"glove"
	DocumentLSTMEmbeddings	hidden_size=512, reproject_word_dimension=256
Humor detection in Spanish	ELMoEmbeddings	"original," "all"
	TransformerWordEmbeddings	"dccuchile/bert-base-spanish-wwm-uncased," layer="-1,-2,-3,-4," layer_mean=True
	FlairEmbeddings	"es-forward"
	WordEmbeddings	"glove"
	DocumentLSTMEmbeddings	hidden_size=512, reproject_word_dimension=256

Table 7 Result of our system on humor, humor controversy, and irony detection tasks (Micro Avg. F1 score, accuracy, recall, and precision; higher is better)

Task	F1 score	Accuracy	Recall	Precision
SemEval-2021 task 7 (humor detection)	0.936	0.922	0.938	0.935
SemEval-2021 task 7 (humor controversy detection)	0.548	0.549	0.602	0.503
SemEval-2018 task 3 (irony detection)	0.712	0.714	0.894	0.593
IberLEF-2021 HAHA task (humor detection in Spanish)	0.877	0.878	0.868	0.887
IberLEF-2019 HAHA task (humor detection in Spanish)	0.807	0.849	0.808	0.807
IberEval-2018 HAHA task (humor detection in Spanish)	0.820	0.866	0.814	0.827

consider 3e-5 and 4e-5 in learning rate, mini-batch size 16 and 32, anneal factor 0.5, 0.6, 0.7, and 0.8, max epochs 2, 3, 6, 10, 15, 20, 35, and 50. For humor and humor controversy detection, we affixed our learning rate 4e-5, mini-batch size 16, anneal factor 0.8, and max epoch 6. Humor detection in Spanish acceded with the same hyper-parameter layouts except for learning rate, anneal factor, and max epochs that are 3e-5, 0.5, and 2, respectively. On the other hand, irony detection occupied the same hyper-parameter configuration except for learning rate 3e-5 and max epochs 50. Finally, we gained our best parameter settings that help to obtain a competitive performance on all of the experimented datasets.

4.4 Results and analysis

In this section, we evaluate the performance of our proposed StackedEmbedding_LA system following the evaluation criteria discussed in Sect. 4.2. We consider the F1 score as the primary evaluation measure along with other standard evaluation measures including recall, precision, and accuracy. We present the result of our method based on test data of individual tasks in Table 7. It shows that our system achieved a reasonably good score in humor and irony detection tasks for both English and Spanish datasets. However, due to the highly twisted nature of the humor controversy detection task, our system obtained a comparatively lower score for

Table 8 Performance analysis of individual models used in our StackedEmbedding_LA system on humor, humor controversy, and irony detection tasks (Micro Avg. F1 score, accuracy, recall, and precision; higher is better)

Category	Model	F1 score	Accuracy	Recall	Precision
Humor detection	StackedEmbedding_LA	**0.936**	**0.922**	**0.938**	0.935
	Performance of individual model on SemEval-2021 task 7				
	GloVe	0.838	0.801	0.840	0.836
	Flair news-forward	0.868	0.838	0.871	0.866
	ELMo	0.881	0.856	0.868	0.894
	BERT	0.906	0.889	0.874	**0.940**
Humor controversy detection	StackedEmbedding_LA	**0.548**	**0.549**	**0.602**	0.503
	Performance of individual model on SemEval-2021 task 7				
	GloVe	0.316	0.425	0.216	0.591
	Flairnews-forward	0.446	0.473	0.344	0.632
	ELMo	0.267	0.448	0.164	0.726
	BERT	0.397	0.499	0.268	**0.763**
Irony detection	StackedEmbedding_LA	**0.712**	**0.714**	**0.893**	0.592
	Performance of individual model on SemEval-2018 task 3				
	GloVe	0.642	0.681	0.723	0.578
	Flair news-forward	0.642	0.660	0.768	0.552
	ELMo	0.629	0.639	0.771	0.531
	BERT	0.679	0.702	0.794	**0.593**
Humor detection in Spanish	StackedEmbedding_LA	**0.877**	**0.878**	**0.867**	0.887
	Performance of individual model on IberLEF-2021 HAHA task				
	GloVe	0.768	0.786	0.709	0.838
	Flair es-forward	0.791	0.812	0.711	0.890
	ELMo	0.796	0.806	0.756	0.840
	BERT	0.863	0.868	0.829	**0.899**

The best results are highlighted in boldface

Table 9 Comparative performance with the systems of other participants' on humor, humor controversy, and irony detection tasks in English

Category	Proposed method	F1 score	Accuracy
Humor detection	*Comparative performance on SemEval-2021 task 7*		
	RoBERTa+ALBERT+Data Augmentation+Adversarial Training [DeepBlueAI] (Song et al. 2021)	0.967	0.960
	RoBERTa-base+RoBERTa-large [SarcasmDET] (Faraj and Abdullah 2021)	0.967	0.960
	StackedEmbedding_LA	**0.936**	**0.922**
	ALBERT+BiLSTM [ZYJ] (Zhao and Tao 2021)	0.934	0.921
	Linear layers+BiLSTM layers+BERT [Team_KGP] (Mondal and Sharma 2021)	0.923	0.903
	BERT+CNN [Tsia] (Guan and Zhou 2021)	0.920	0.896
Humor controversy detection	*Comparative performance on SemEval-2021 task 7*		
	RoBERTa+ALBERT+Data Augmentation+Adversarial Training [DeepBlueAI] (Song et al. 2021)	0.625	0.465
	RoBERTa-base+RoBERTa-large+BERT-base+BERT-large [SarcasmDET] (Faraj and Abdullah 2021)	0.627	0.469
	StackedEmbedding_LA	**0.548**	**0.549**
	ALBERT+BiLSTM [ZYJ] (Zhao and Tao 2021)	0.460	0.440
	Multitask learning+Ensemble of pre-trained models [Amherst685] (Zylich et al. 2021)	0.484	0.522
	SVM+Lightweight features [RedwoodNLP] (Chi and Chi 2021)	0.488	0.502
Irony detection	*Comparative performance on SemEval-2018 task 3*		
	StackedEmbedding_LA	**0.712**	**0.714**
	Densly connected LSTM with multitask learning [THU_NGN] (Wu et al. 2018)	0.705	0.735
	Word and character-based bidirectional LSTM [NTUA-SLP] (Baziotis et al. 2018)	0.672	0.732
	LR+SVM with various embedding, word-based and handcrafted features [WLV] (Rohanian et al. 2018)	0.650	0.643

The performances of our system are highlighted in boldface

this task. Comparative results presented in Table 9 demonstrated that other state-of-the-art systems also suffered in this task. In the future, we have a plan to incorporate some domain-specific technologies to tackle the challenges of this task.

In our StackedEmbedding_LA model, we have stacked different word embedding models including GloVe, ELMo, BERT, and Flair news-forward. To compare the result of our proposed StackedEmbedding_LA with individual components, we have evaluated the performance of each component on all mentioned tasks. We place the results of these experiments on each task in Table 8. Here, we can observe that the performance of individual models is significantly lower than the performance of StackedEmbedding_LA in each task. Experimental results show that our proposed method outperforms individual models by at least 3%, and at best 9.8% in the humor detection task, in terms of the primary evaluation metric F1 score. In the humor controversy detection task the performance difference is at least 10.2%, and at best 28.1%. In a similar trend, the minimum and the maximum difference in F1 score in the irony detection task are 3.3%, and 8.3%, respectively. And, in the Spanish humor detection task, our proposed model outperforms individual models by at least 1.4%, and at best 10.9%.

However, all these mentioned components showed a similar performance individually. The BERT model obtained the best result compared to other models in discrete performances. ELMo attained second-best completion while detecting humor in English, but it does not perform well in the other three tasks. Flair secured good results in humor controversy detection and humor detection in Spanish. GloVe performed best in the irony detection task and fails to obtain comparatively good results in other tasks. In contrast, when we combined these models and performed the stacked embeddings approach, we achieved better performance in all these tasks compared to individual models' performance. Moreover, to seize the complex relationships among sentences, we applied document LSTM embeddings. We divided feature vectors into four feature levels and utilized the last feature zone that affirms StackedEmbedding_LA as a more proficient system.

4.5 Comparison with related work

To authenticate the efficacy of our proposed StackedEmbedding_LA method, we compared the performance of our system with some other submitted approaches on both humor and irony detection shared tasks. The comparative

Table 10 Comparative performance with the systems of other participants' on humor detection task in Spanish

Proposed method	F1 score	Accuracy
Comparative performance on IberLEF-2021 HAHA task		
BERT-base multilingual cased+BETO+ALBERT+ Sentiment analysis BERT+RoBERTa [Jocoso] (Grover and Goel 2021)	0.885	0.889
StackedEmbedding_LA	**0.877**	**0.878**
ColBERT+BETO [ColBERT] (Annamoradnejad and Zoghi 2021)	0.869	–
BERT-base multilingual uncased+LSTM [kuiyongyi] (Kui 2021)	0.868	–
BERT-base multilingual [TECHSSN] (Nanda et al. 2021)	0.767	0.797
Comparative performance on IberLEF-2019 HAHA task		
BERT-base multilingual cased+Fastai+Multinomial Naïve Bayes+TF-IDF+Logistic regression [adilism] (Ismailov 2019)	0.821	0.855
ULMFiT+Fastai+Byte pair encoding [bfarzin] (Farzin et al. 2019)	0.810	0.846
StackedEmbedding_LA	**0.807**	**0.849**
μTC+Sparse and dense word representations+Linear SVM [INGEOTEC] (Ortiz-Bejar et al. 2019)	0.788	0.828
BERT-base multilingual cased [BLAIR_GMU] (Mao and Liu 2019)	0.784	0.822
Lemmatization+Spanish word embeddings+Handcrafted features+BiGRU [UO_UPV2] (Ortega-Bueno et al. 2019)	0.773	0.824
Comparative performance on IberEval-2018 HAHA task		
StackedEmbedding_LA	**0.820**	**0.866**
EvoMSA+EvoDAG [INGEOTEC] (Ortiz-Bejar et al. 2018)	0.797	0.845
BiLSTM+Linguistically motivated features+Attention-based Word2Vec models [UO_UPV] (Ortega-Bueno et al. 2018)	0.785	0.845
SVM+Bag of character n-grams [ELiRF-UPV] (Castro et al. 2018)	0.772	0.836

The performances of our system are highlighted in boldface

performance of our method based on test data against the other participants' systems in individual shared tasks are presented in Table 9 and Table 10.

At the SemEval-2021 task 7 (Meaney et al. 2021) humor and humor controversy detection task, DeepBlueAI (Song et al. 2021) proposed a system where they ensemble predictions from a RoBERTa (Liu et al. 2019b) and an ALBERT (Lan et al. 2019) model. Besides fine-tuning, they augmented datasets through pseudo-labeling and added these augmented test sets with training data. For improving generalization, they utilized adversarial training (Miyato et al. 2016) by adding perturbations to the embedding layer. Later implementing multisample dropout, final predictions were gained. In score comparison against our proposed StackedEmbedding_LA method, the F1 score and accuracy of the mentioned team preceded by 3.1% and 3.8% in humor detection, whereas the F1 score preceded by 7.7% and accuracy preceded by 8.4% in humor controversy detection. Doing a constructive analysis, we find some lackings in that system. First of all, text embeddings may be extremely sensitive to even minor changes. In reality, a minor perturbation may cause a sentence to have an inaccurate syntactic structure or entirely different semantic meaning, leading to a difference between adversarial correctness, which is defined as robustness, and standard accuracy, which is defined as generalization. Models trained with an adversarial purpose frequently exhibit an increase in the robust accuracy but a drop in the standard accuracy because the features learned by the robust and standard classifiers can be fundamentally different (Tsipras et al. 2018; Raghunathan et al. 2019). SarcasmDET (Faraj and Abdullah 2021) also utilized an ensemble-based system where they aggregated RoBERTa-base and large models for humor detection and included BERT-base and large models in their ensemble-based system for humor controversy detection. However, the combination of heavyweight features resulting in larger network weights may cause higher generalization errors. Besides, similar kinds of textual context may be extracted by employing two similar kinds of transformer models thus lack of contextual diversity. ZYJ (Zhao and Tao 2021) designed their system through capturing input embedding, position encoding, and token-type embedding in ALBERT in which the last hidden layers are added with the BiLSTM for obtaining the feature vector. The later feature vector is concatenated with the original output of ALBERT. Team_KGP (Mondal and Sharma 2021) proposed a system based on two different types of fine-tuning methods by using linear layers and BiLSTM layers on top of the pre-trained BERT model. Tsia (Guan and Zhou 2021) combined BERT with CNN architecture finding an average result in humor detection. Amherst685 (Zylich et al. 2021) employed their method utilizing multitask learning and ensembling of different pre-trained language models for detecting controversial humor, whereas RedwoodNLP (Chi and Chi 2021) implied SVM utilizing lightweight features as inputs. Here, we observe the lacking of utilizing different types of transformer models in these proposed systems

that limited their ability to capture the diverse contextual information from the input. In this viewpoint, we can find a very strong coverage through stacking various word embedding models.

Diverting in the multilingual field, at the IberLEF-2021 HAHA (Chiruzzo et al. 2021) Spanish humor detection task, participants mostly used transformer models trained with Spanish text data. Jocoso (Grover and Goel 2021) conducted a hard voting-based system aggregating five models including BERT-base multilingual cased (Devlin et al. 2018), BETO (Canete et al. 2020), ALBERT (Lan et al. 2019), sentiment analysis BERT (de Arriba Serra et al. 2021), and RoBERTa (Liu et al. 2019b) later employing multinomial naïve Bayes classifier with TF-IDF features. Their proposed system was preceded by 0.8% and 1.1% in F1 score and accuracy, respectively. For achieving better performance, they conducted a training process on five individual models separately. In contrast, we stacked various word embedding models and utilized document LSTM embeddings to capture both the word and sentence context in a unified model and obtained almost similar kinds of scores as the top performing system Jocoso. ColBERT (Annamoradnejad and Zoghi 2021) acclimated ColBERT (Annamoradnejad and Zoghi 2020) to Spanish utilizing BETO-uncased (Canete et al. 2020). They mainly focused on separating different neural paths where BETO embeddings are fed into parallel hidden layers to extract latent features from the input text. Later, three sequential layers are placed on top of the exploited hidden layers to forecast the final output. Kuiyongyi (Kui 2021) unified a fine-tuned BERT-base multilingual uncased (Devlin et al. 2018) model with an LSTM network and employed some data cleaning methods including repeated characters or words replacement, emoticons, and HTML tags cleaning. TECHSSN (Nanda et al. 2021) also mobilized parallel hidden layers of the neural network to capture the inner contents of each sentence as well as the whole text. But the difference is that they applied BERT-base multilingual model (Devlin et al. 2018) with some preprocessing techniques including stemming and lemmatizing.

At the IberLEF-2019 HAHA (Chiruzzo et al. 2019) Spanish humor detection task, adilism (Ismailov 2019) exploited fine-tuned BERT-base multilingual cased (Devlin et al. 2019) model with the fastai library (Howard and Gugger 2020). In the BERT portion, they utilized the output of the last layer for the [CLS] token with a tanh activation including a linear layer, a dropout layer, and another linear layer with a binary cross-entropy loss. They also conducted experiments on multinomial naïve Bayes (Wang and Manning 2012) with unigram and bigram TF-IDF features and logistic regression to procure the final predictions. However, only a single BERT-base multilingual model may not capture the word contexts from diverse perspectives and multinomial naïve Bayes focus on the tag of the texts for final predictions rather analyzing the latent features of the text. In contrast, our mentioned stacked embeddings module can comfortably alleviate these issues. Besides, logistic regression on top of the learning model can not handle complex associations in the text as it can not afford the repeated data. In humor and irony detection, we have to depend on multiple observations on repeated data, in such cases, this method can be treated as a vulnerable system. For mitigating this problem, we apply document LSTM embeddings that can easily grip the complex relationships among the sentences. Through these comparisons, we can validate the robustness of our StackedEmbedding_LA system. Bfarzin (Farzin et al. 2019) utilized universal language model fine-tuning (ULMFiT) (Howard and Ruder 2018) with fastai library (Howard and Gugger 2020) where they perform tokenization with Byte Pair Encoding (BPE)(Sennrich et al. 2015). However, in their proposed system, they executed the fine-tuning process of the single language model thus lack of diversity, whereas we employ multiple fine-tuned embeddings through stacking in a unified model to make our model robust for capturing diverse contextual features. INGEOTEC (Ortiz-Bejar et al. 2019) employed μTC (an automated text categorization framework) (Tellez et al. 2018) with sparse and dense word representations where linear SVM was reported as their best performing system. This method did not detain the miscellaneous context of words and grasp the compositionality of words. Moreover, they also applied fastText (Bojanowski et al. 2017) and Flair (Akbik et al. 2018) including the various amalgamation of token embeddings that range from simple characters to BERT (Devlin et al. 2019), B4MSA (Tellez et al. 2017), and EvoMSA (Graff et al. 2020) but did not procure any enhancement. In our proposed method, we exploit stacking of different word embeddings with document LSTM embeddings which not only depicts a better delineation of the diverse context of words but also represents the whole sentence context effectually. However, the selective feature level included a new dimension on the overall performance of our proposed system. BLAIR_GMU (Mao and Liu 2019) utilized BERT-base multilingual cased (Devlin et al. 2019) where they considered the last layer output corresponding to the [CLS] token and included a linear output for predicting the Spanish humor labels. This method did not procure word contexts from various aspects. On the contrary, UO_UPV2 (Ortega-Bueno et al. 2019) employed lemmatization utilizing FreeLing (Padró and Stanilovsky 2012), an in-house produced collection of Spanish word embeddings and handcrafted features including structural and content, stylistic and affective based on LIWC (Pennebaker et al. 2001). Later, they applied produced vector as the initial hidden state of attention-based BiGRU (Chung et al. 2014) succeeded by three dense layers. Their applied handcrafted features are confined to some specific word categories as well as extracted only from the extrinsic context

of words. Moreover, BiGRU did not perform well on the dataset utilizing longer sequences where the attention process escalated the training time. Our proposed method attenuates these issues by employing the stacked version of word embeddings that captures the context of words from different viewpoints as well as document LSTM embeddings functions well on longer sequences of input data. Moreover, we segment feature vectors into four feature levels and exploit the last feature zone to procure the final prediction labels that validate our proposed system computationally more competent.

At the IberEval-2018 HAHA task (Castro et al. 2018) Spanish humor detection, INGEOTEC (Ortiz-Bejar et al. 2018) applied μTC (Tellez et al. 2018) including naïve Bayes where μTC (an automated text categorization framework) procured text models enhancing a performance measurement and utilized an SVM with a linear kernel classifier. Besides mentioned team implied another tool named EvoMSA (Graff et al. 2020) that applied the EvoDAG (Graff et al. 2016) classifier. EvoDAG is a genetic programming system with tournament selection that performs genetic operations from the root motivated by geometric semantic crossover. Though these tools are delineated for defining a parameter space illustrating a huge number of text classifiers (μTC) as well as reliable for the problems of unbalanced classes (EvoMSA), they can not verify the diversity of word contexts and learn the compositionality of words. For detaching these lackings, we employ the stacking of various word embeddings and merge these stacked embeddings into a document LSTM embeddings for representing the whole sentence context constructively. Besides, our proposed method outperforms this top performing system by 2.32% and 2.1% in terms of evaluation metrics F1 score and accuracy, respectively, which authenticates the effectiveness of our method. UO_UPV (Ortega-Bueno et al. 2018) utilized BiLSTM (Graves and Schmidhuber 2005) and a bunch of linguistically motivated features including structural and content, stylistic, and affective features. ELiRF-UPV (Castro et al. 2018) proposed two systems where the first system is composed of SVM and a bag of character n-grams and the second system is integrated with CNN (Albawi et al. 2017). SVM is not suited for large and noisy datasets, we can handle this matter by exploiting a diverse set of word embeddings that can extract words from hidden contexts more effectually than the classifier. Moreover, CNN conducts convolutional layers and maximum pooling layers to extract features thus it delineates a better representation of image than text. For mitigating this issue, we employ document LSTM embeddings that can seize long-term dependencies between word sequences thus depicting a finer rendition of the text.

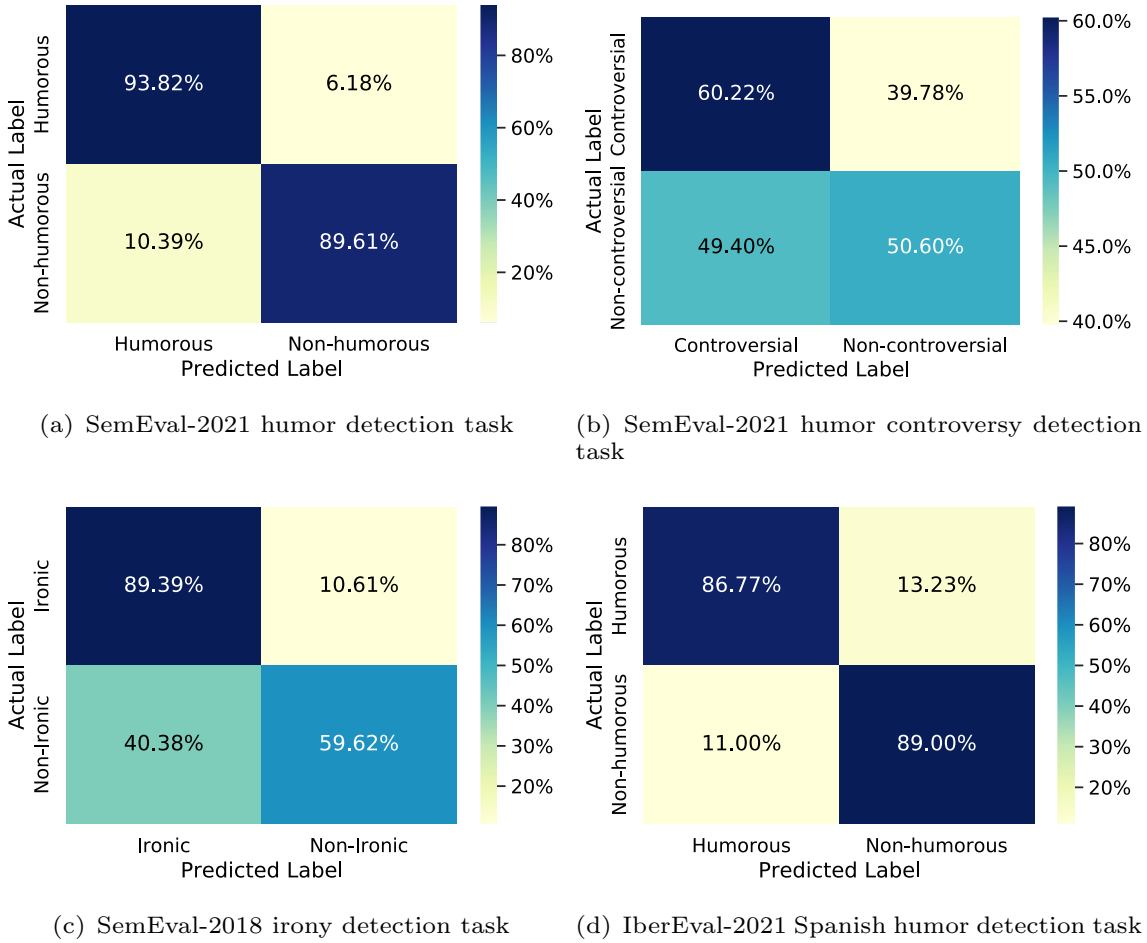
At SemEval-2018 task 3 (Van Hee et al. 2018) irony detection, THU_NGN (Wu et al. 2018) proposed a densely connected LSTM network-based system in combination with a multitask learning strategy. There is inadequacy in the ensemble of preprocessing techniques as well as utilizing fine-tuned word embedding models. Our proposed system outperforms this top performing system by 0.7% in terms of the primary evaluation metric F1 score. NTUA-SLP (Baziotis et al. 2018) prioritized majority voting where two independent models are utilized which are based on word and character-based BiLSTM for capturing both syntactic and semantic context of tweets. Here, majority voting implies the model expansion where two individual model outcomes are combined for predicting the labels. WLV (Rohanian et al. 2018) submitted a model which is also an ensemble voting-based approach where logistic regression and SVM are applied including various embedding, word-based, and handcrafted features.

Analyzing these state-of-the-art methods, our proposed system is procured a firm position through stacking diverse word embeddings including GloVe, ELMo, Flair, and BERT. GloVe vectorizes the text from both global and local perspectives, Flair captures the different meanings of a certain word, ELMo extracts the latent context of the word, whereas BERT seizes specific word context based on overall word contents. This integrated module scrutinizes the text not only from a single point of view but also from multiple aspects. We fine-tune all of these word embedding models and exploit fine-tuned Spanish BERT and Spanish Flair models for gaining the label of Spanish humor that helps us to secure a good result in the multilingual field also. Later, we conduct these stacked word embedding modules through document LSTM embeddings for representing the whole sentence context. Another benefaction is segmenting feature vectors into the multilayer frame. Through splitting lightweight feature vectors, we can easily learn useful features representation of the text as well as memory efficiency. Moreover, lightweight features give a better delineation of the sentence, capture the sentence context more effectually than heavyweight features, and establish a memory efficient as well as robust system (Zhang et al. 2016; Tay et al. 2019; Desai et al. 2020). Therefore, in the emerging multilingual field, our proposed system secured a strong place and competitive performance in both humor and irony detection compared to other state-of-the-art system.

5 Discussion

5.1 Error analysis

To perform the error analysis, we analyze the performance of our StackedEmbedding_LA system with the confusion matrix depicted in Fig. 2. In the confusion matrix, we include humor and humor controversy detection from SemEval-2021, irony detection from SemEval-2018, and Spanish



(a) SemEval-2021 humor detection task

(b) SemEval-2021 humor controversy detection task

(c) SemEval-2018 irony detection task

(d) IberEval-2021 Spanish humor detection task

Fig. 2 Prediction synopsis of our system. **a** Confusion matrix of humorous text classification. **b** Confusion matrix of humorous controversial text classification. **c** Confusion matrix of ironic text classification. **d** Confusion matrix of Spanish humorous text classification

humor detection from IberEval-2021 shared tasks. We observe that 6.18% of texts are mislabeled as non-humorous and 10.39% of texts are misidentified as humorous in the humor identification task. On the contrary, 49.40% of non-controversial texts and 39.78% of controversial texts are misapprehended as controversial and non-controversial humorous texts, respectively. In the case of humor detection in Spanish, the misidentification rate of humorous texts is 13.23% which is higher than the misidentification rate of non-humorous texts. Besides, 10.61% of ironic tweets and 40.38% of non-ironic tweets are incorrectly interpreted as non-ironic and ironic tweets consecutively. The rate of misclassified non-humor and non-ironic texts is higher which indicates our system lags in classifying non-humor and non-ironic texts appropriately. For controversial humorous text detection, the ratio of negative categorization is relatively higher than the other tasks.

Besides, we conduct a study on misclassified texts by our system to address the reasons for erroneous predictions, some of which are illustrated in Table 11. It shows that the first example in the humor detection task is ambiguous and the second example is too small to understand its context. It indicates that ambiguity and shortness in texts limit the performance of our system. Furthermore, the system fails to capture the context of long text with too much irrelevant information. The presence of native and multilingual word context, very short erratic word form, immoderate, extended, and redundant use of emojis and hashtags lessen the potentiality of our system to distill the right interpretation. Applying a proper strategy to handle these issues might be fruitful in excel the performance of our system.

5.2 Qualitative analysis

To perform the qualitative analysis of our proposed StackedEmbedding_LA method, we have compared the prediction outcome of the individual embedding models for the few test input. The comparison is illustrated in Table 12. Here, we observe that for these articulated samples we get wrong predictions while using individual embedding models

Table 11 Examples of misclassified texts

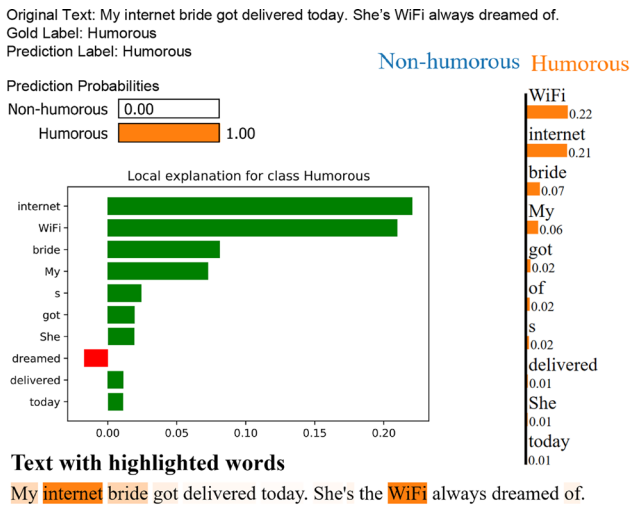
Task	Text	Predicted label	True label
Humor detection	*SemEval-2021 task 7*		
	#1: Why does today feel like a Sunday!	Humorous	Non-humorous
	#2: And then there's my dad..😳	Non-humorous	Humorous
	#3: If you want to be a General Motors engineer, your memory needs to be perfect. You have to recall everything	Non-humorous	Humorous
Humor controversy detection	*SemEval-2021 task 7*		
	#1: If you want to be a General Motors engineer, your memory needs to be perfect. You have to recall everything	Non-controversial	Controversial
	#2: What do you call a black man on the moon? An astronaut	Non-controversial	Controversial
	#3: Whats the hardest part of a vegetable to eat? The wheelchair	Controversial	Non-controversial
Irony detection	*SemEval-2018 task 3*		
	#1: I can't even begin to explain my frustration	Ironic	Non-ironic
	#2: What a lovely day to drive #not #boo #fog #work 😳 😳 😳	Non-ironic	Ironic
	#3: #mcfc trying to kill their opponents	Ironic	Non-ironic
Humor detection in Spanish	*IberLEF-2021 HAHA Task*		
	#1: Para que te salgan mariposas en el estómago primero hay que pasar un tiempo con capullos.	Humorous	Non-humorous
	#2: ¿Sabéis esa gente que tiene un morro que se lo pisa y se apunta a un bombardeo si es gratis...? Pues ahora se llaman allegados.	Non-humorous	Humorous
	#3: Recién cambié un enchufe en dos patadas.	Humorous	Non-humorous

Table 12 Qualitative analysis of individual models of StackedEmbedding_LA system on humor and humor controversy detection task

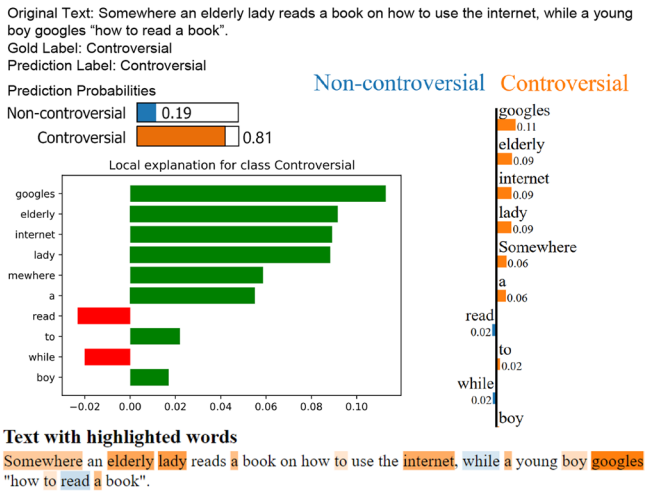
Text	Stacked Embedding _LA	GloVe	Flair	ELMo	BERT	Gold label
Humor detection						
If alcohol influences short-term memory, what does alcohol do?	1	0	0	0	0	1
hey, brands. you don't need to do a 9/11 post. it's ok. take the day off. we will pick back up with the pillsbury doughboy tomorrow. whether chester cheeto is for or against terrorism is not in question	1	0	0	0	0	1
2020 is pretty cool because every once in a while you get to experience every emotion at once. And it's just like ahhh omg please stop hahah	1	0	0	0	0	1
Humor controversy detection						
It's Friday night and I'm out of control! Getting a bit wild tonight cuz I'm about to put on my good pajamas, and eat some Froot Loops on the couch with a fluffy blanket	1	0	0	0	0	1
I burned a kid in a wheelchair today. Hot wheels	1	0	0	0	0	1
I made it halfway to Mexico before I realized that those sirens were just coming from the song on my radio	1	0	0	0	0	1

including GloVe, BERT, ELMo, and Flair. But when we combine these embedding models in our unified StackedEmbedding_LA model, we get the correct prediction labels. The GloVe embeddings model vectorizes the text from both global and local perspectives. It focuses on the words' co-occurrences over the whole corpus (Pennington et al. 2014). For this reason, it can not predict the diversity of the contextual word. The Flair model catches the multifarious meanings of a certain word but it does not focus on word similarity and co-occurrences (Akbik et al. 2018). ELMo is a deep contextualized word representation that models convoluted characteristics of word use. But it does not focus on texts from a global or local perspective (Peters et al. 2018). BERT model represents word embeddings based on the context of the word but it is lack of ability to handle long text sequences.

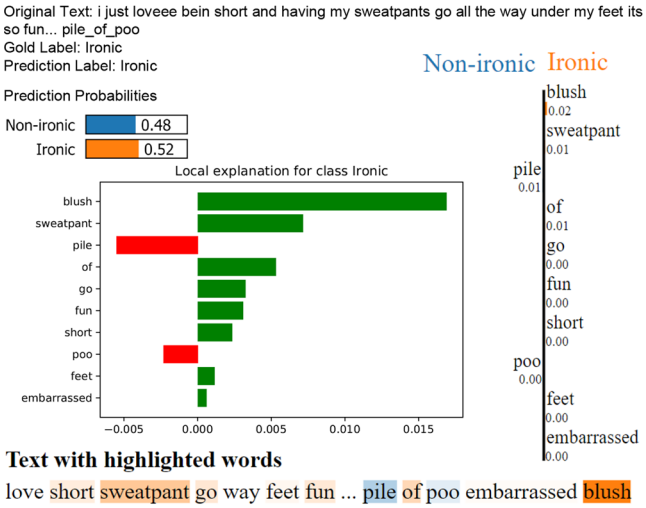
When we stacked all these four embedding models, we can overcome the limitations of individual models by combining all the features of these embedding models. Our StackedEmbedding_LA system can capture the diversity of



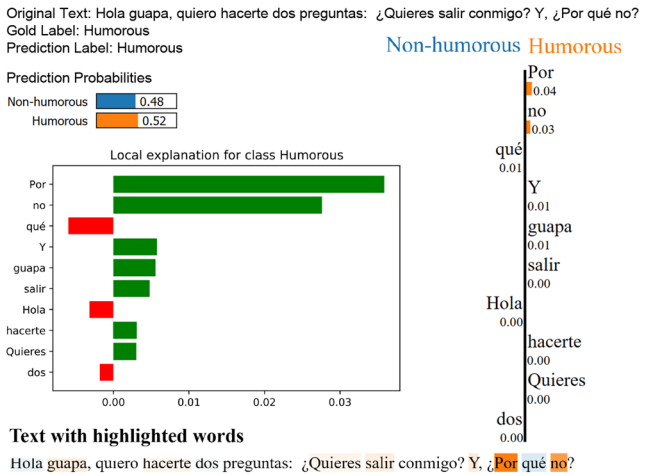
(a) SemEval-2021 humor detection task



(b) SemEval-2021 humor controversy detection task



(c) SemEval-2018 irony detection task



(d) IberEval-2021 Spanish humor detection task

Fig. 3 LIME explanation for the prediction of our system. **a** Explanation of humorous text classification. **b** Explanation of humorous controversial text classification. **c** Explanation of ironic text classification. **d** Explanation of Spanish humorous text classification

contextual sentences and predict the true label where the individual model fails.

5.3 Explainability of our proposed model

We have utilized local interpretable model-agnostic explanations (LIME) to explain the prediction of our model in an interpretable manner. LIME provides us a qualitative insight presenting textual or visual artifacts of the relationship between the components of an instance (e.g., words in the text) and the response of a classifier.

LIME: (Ribeiro et al. 2016) suggested a feature-based approach called "LIME" makes any classifier interpretable by approximating it locally. They leverage a perturbation-based strategy in which they randomly adjust a tiny portion of the input and then assess the impact on the model output to explain the prediction of a classifier (Alzubaidi et al. 2021). For a single data sample p′, LIME generates a perturbed dataset by eliminating a random subset of the instance's words. These perturbed samples are fed to our classifier to see how it would predict them. For text data, LIME treats the presence and absence of each word as a feature. The absence of certain word or words in a new sample influences the predicted label for the sample and the confidence score. Considering the effect, LIME weights the samples in the resulting dataset following their proximity to

p′. Samples close to p′ are given a large weight and samples far away from p′ are given a small weight. In our experiment with LIME, we create 5000 perturbed samples from a single data sample to analyze the prediction of our model and the influence of words.

The LIME visualization of various examples of our tasks is depicted in Fig. 3. LIME provides individual feature relevance and high feature contribution highlighting the top words which significantly influence the system to make the classification choice. Figure 3(a) describes the LIME explanation of a humorous text being predicted as humourous by our system StackedEmbedding_LA. Here, the orange color and the blue color represent classes humorous and non-humorous, respectively. From the human perspective, the humor in the sentence lies in the thought that the speaker has purchased a "wife" who is a WiFi device. In this case, the wordplay "internet bride" and the analogy between a WiFi device and a loving partner are humorous. As seen in the LIME visualization of this text, the words "Internet," "WiFi," and "bride" have the highest feature weights. It represents the fact that these words are most impactful for the sentence to predict as humorous which is consistent with the explanation from a human standpoint. The illustration of the LIME explanation for the multilingual humor is depicted in Fig. 3(d) where the sentence is in Spanish. Here, "Hola guapa, quiero hacerte dos preguntas: ¿Quieres salir conmigo? Y, ¿Por qué no?" that translates into "Hello beautiful, I want to ask you two questions: Do you want to go out with me? And why not?" in English using Google Translate. We can observe that this sentence is predicted with 52% probability in the humorous category where "¿Por (By)" occupies the maximum feature weight to label this sentence as humorous. Moreover, "no (No)," "Y (And)," "guapa (pretty)," "salir (Go out)," "hacerte (Make you)," and "¿Quieres (Want)" also contribute to the humorous category. The abrupt change in attitude from self-assured to self-deprecating is what makes this sentence funny. The questioner appears to be willing to take a chance even though they anticipate receiving a "no" in response. An upbeat and playful tone is produced by the union of self-assurance and self-awareness.

The next example in the Fig. 3(b) is predicted as controversial humor with a 81% probability. Here, the controversial humor class is represented by the orange color and the non-controversial humor class is represented by the blue color. When it comes to technology and reading, this instance illustrates the varying attitudes and skills of the various generations. This observation, which is meant to be humorous, might be seen by others as a generalization, and therefore, it is controversial. Our system concentrates on the terms "googles," "elderly," "internet," and "lady" to predict the statement as controversial according to LIME representation which indicates our proposed model emphasizes the appropriate terms to comprehend the context of the sentence. Figure 3(c) depicts the LIME visualization for irony detection task. Here, the orange color and the blue color represent the ironic and non-ironic classes, respectively. The speaker claims that they "simply love being short" which contradicts the fact that for being short it is difficult to find clothes that fit properly, like sweatpants. In this context, our system provides the most importance to the words "blush," "sweatpant," and "embarrassed" to predict the sentence as ironic.

6 Conclusion and future directions

In this paper, we have employed a stacked embeddings model where we aggregated numerous word embedding models including GloVe, ELMo, Flair, and BERT for extracting the diverse context of the word. GloVe captured the context of the word both from global and local perspectives, ELMo distilled the hidden context of the word, Flair verified different meanings of a single word, whereas BERT extracted certain word contexts following other inclusive word contents. Moreover, we conducted Spanish BERT and Spanish Flair word embedding models for procuring the labels of Spanish humor. We fine-tuned all of these word embedding models that capture the information of the word more effectually from semantical and contextual viewpoints. We also employed document LSTM embeddings on these stacked word embedding feature vectors for capturing the context of the whole sentence. Utilizing multiple feed-forward linear architectures, we segmented merged feature vectors into four feature levels and implemented the last feature frame for obtaining the final prediction labels. We conducted a small portion of feature vectors as it resulted in smaller network weights, ensured a more robust network, and lower generalization error. Experimental results depicted that our method surpassed the highest score of the contest in the SemEval-2018 irony detection and IberEval-2018 Spanish humor detection tasks, whereas delivered a competitive result in the SemEval-2021 humor and humor controversy detection, IberLEF-2019, and IberLEF-2021 Spanish humor detection tasks. Analyzing our proposed system, it is visible that stacking is a better ensemble aggregation method in comparison with other state-of-the-art ensembling strategies. Besides, it has achieved a good performance in multilingual datasets that contains different traits.

In the future, we have a plan to employ topic modeling capturing the arrangement of word bunches and recurrences of words in the text for diminishing the process complexity of our proposed system. Besides, we intend to utilize more lightweight feature vectors and reduce computational resources while retaining the high accuracy of our proposed method. Both lightweight features and low computational resources occupy small portions of weights in the network

which procures high memory efficiency and less processing time. Moreover, we will validate our system on other multilingual datasets for ensuring its robustness and portability. Exploiting all these forthcoming steps, we hope our proposed system will detach its lacking and capitulate to the highest performance.

Author Contributions Radiathun Tasnia, Nabila Ayman, and Afrin Sultana were responsible for conceptualization, methodology, software, investigation, writing the original draft, and reviewing. Abu Nowshed Chy was involved in conceptualization, methodology, validation, writing—reviewing and editing, and supervision. Masaki Aono contributed to conceptualization, methodology, writing—review and editing, and supervision.

Funding Not applicable.

Availability of data The datasets used within this study are publicly available and respective references to these datasets are included in the manuscript.

Declarations

Conflict of interest The authors declare no competing interests.

References

Akbik A, Blythe D, Vollgraf R (2018) Contextual string embeddings for sequence labeling. In: Proceedings of the 27th international conference on computational linguistics, pp 1638–1649. https://aclanthology.org/C18-1139/

Albawi S, Mohammed TA, Al-Zawi S (2017) Understanding of a convolutional neural network. In: 2017 International conference on engineering and technology (ICET), IEEE, pp 1–6. https://doi.org/10.1109/ICEngTechnol.2017.8308186

Altin LSM, Bravo À, Saggion H (2019) Lastus/taln at haha: Humor analysis based on human annotation. In: IberLEF@ SEPLN, pp 145–150

Alzubaidi L, Zhang J, Humaidi AJ et al (2021) Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. J Big Data 8:1–74. https://doi.org/10.1186/s40537-021-00444-8

Amir S, Wallace BC, Lyu H, et al (2016) Modelling context with user embeddings for sarcasm detection in social media. arXiv preprint arXiv:1607.00976https://doi.org/10.48550/arXiv.1607.00976

Annamoradnejad I, Zoghi G (2020) Colbert: Using bert sentence embedding for humor detection. arXiv preprint arXiv:2004.12765https://doi.org/10.48550/arXiv.2004.12765

Annamoradnejad I, Zoghi G (2021) Colbert at haha 2021: Parallel neural networks for rating humor in spanish tweets. In: Proceedings of the Iberian languages evaluation forum (IberLEF 2021), CEUR workshop proceedings, Málaga, Spain

Barbieri F, Saggion H (2014) Modelling irony in twitter. In: Proceedings of the student research workshop at the 14th conference of the european chapter of the association for computational linguistics, pp 56–64. https://doi.org/10.3115/v1/e14-3007

Baziotis C, Pelekis N, Doulkeridis C (2017) Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In: Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017), pp 747–754. https://doi.org/10.18653/v1/S17-2126

Baziotis C, Athanasiou N, Papalampidi P, et al (2018) Ntua-slp at semeval-2018 task 3: Tracking ironic tweets using ensembles of word and character level attentive rnns. arXiv preprint arXiv:1804.06659https://doi.org/10.48550/arXiv.1804.06659

Bharti SK, Babu KS, Jena SK (2015) Parsing-based sarcasm sentiment recognition in twitter data. In: 2015 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), IEEE, pp 1373–1380. https://doi.org/10.1145/2808797.2808910

Bojanowski P, Grave E, Joulin A et al (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguistics 5:135–146. https://doi.org/10.1162/tacl_a_00051

BREDIN H (1997) The semantic structure of verbal irony. J Literary Semantics 26(1):1–20

Brône G (2017) Cognitive linguistics and humor research. In: The Routledge handbook of language and humor. Routledge, pp 250–266

Brône G (2021) The multimodal negotiation of irony and humor in interaction. Figurative Language Intersubject Usage 11:109

Brône G, Feyaerts K, Veale T (2006) Introduction: cognitive linguistic approaches to humor. Humor Int J Humor Res 19(3):203–228

Buschmeier K, Cimiano P, Klinger R (2014) An impact analysis of features in a classification approach to irony detection in product reviews. In: Proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis, pp 42–49. https://doi.org/10.3115/v1/w14-2608

Canete J, Chaperon G, Fuentes R, et al (2020) Spanish pre-trained bert model and evaluation data. Pml4dc at iclr 2020:2020

Castro S, Chiruzzo L, Rosá A (2018) Overview of the haha task: Humor analysis based on human annotation at ibereval 2018. In: IberEval@ SEPLN, pp 187–194

Cattle AG, Zhao Z, Papalexakis EE, et al (2019) Generating document embeddings for humor recognition using tensor decomposition. In: CEUR workshop proceedings, p 151

Cañete J, Chaperon G, Fuentes R, et al (2020) Spanish pre-trained bert model and evaluation data. In: PML4DC at ICLR 2020

Chi N, Chi R (2021) Redwoodnlp at semeval-2021 task 7: Ensembled pretrained and lightweight models for humor detection. In: Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021), pp 1209–1214. https://doi.org/10.18653/v1/2021.semeval-1.171

Chiruzzo L, Castro S, Etcheverry M, et al (2019) Overview of haha at iberlef 2019: Humor analysis based on human annotation. In: IberLEF@ SEPLN, pp 132–144

Chiruzzo L, Castro S, Góngora S, et al (2021) Overview of HAHA at IberLEF 2021: detecting, rating and analyzing humor in Spanish. Procesamiento del Lenguaje Natural 67

Chung J, Gulcehre C, Cho K, et al (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555https://doi.org/10.48550/arXiv.1412.3555

CLARK H (1984) On the pretense theory of irony. J Exp Psychol General 113:121–126

Colston HL, Gibbs RW Jr (1998) Analogy and irony: rebuttal to "rebuttal analogy''. Metaphor Symbol 13(1):69–75

Davidson T, Warmsley D, Macy M, et al (2017) Automated hate speech detection and the problem of offensive language. In: Proceedings of the international AAAI conference on web and social media, pp 512–515. https://doi.org/10.48550/arXiv.1703.04009

de Arriba Serra A, Oriol Hilari M, Franch Gutiérrez J (2021) Applying sentiment analysis on spanish tweets using beto. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021): co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII international conference of the Spanish society for natural language processing: Málaga, Spain, September, 2021, CEUR-WS. org, pp 1–8. http://hdl.handle.net/2117/356656

Desai S, Goh G, Babu A, et al (2020) Lightweight convolutional representations for on-device natural language processing. arXiv preprint arXiv:2002.01535https://doi.org/10.48550/arXiv.2002.01535

Devlin J, Chang MW, Lee K, et al (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805https://doi.org/10.48550/arXiv.1810.04805

Devlin J, Chang MW, Lee K (2019) Kristina, toutanova. Bert: pre-training of deep bidirectional, transformers for language understanding. In: NAACL 2(4):5

Faraj D, Abdullah M (2021) Sarcasmdet at semeval-2021 task 7: Detect humor and offensive based on demographic factors using roberta pre-trained model. In: Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021), pp 527–533. https://doi.org/10.18653/v1/2021.semeval-1.64

Farzin B, Czapla P, Howard J (2019) Applying a pre-trained language model to spanish twitter humor prediction. arXiv preprint arXiv:1907.03187https://doi.org/10.48550/arXiv.1907.03187

Ghosh A, Li G, Veale T, et al (2015) Semeval-2015 task 11: sentiment analysis of figurative language in twitter. In: Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), pp 470–478. https://doi.org/10.18653/v1/s15-2080

Giora R, Fein O (1999) Irony: context and salience. Metaphor Symbol 14(4):241–257

Giudice V (2019) Aspie96 at haha (iberlef 2019): Humor detection in spanish tweets with character-level convolutional rnn. In: IberLEF@ SEPLN, pp 165–171

González JÁ, Hurtado LF, Pla F (2018) Elirf-upv at semeval-2018 tasks 1 and 3: affect and irony detection in tweets. In: Proceedings of The 12th international workshop on semantic evaluation, pp 565–569. https://doi.org/10.18653/v1/s18-1092

González-Ibánez R, Muresan S, Wacholder N (2011) Identifying sarcasm in twitter: a closer look. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, pp 581–586

Graff M, Tellez ES, Miranda-Jiménez S, et al (2016) Evodag: a semantic genetic programming python library. In: 2016 IEEE international autumn meeting on power, electronics and computing (ROPEC), IEEE, pp 1–6. https://doi.org/10.1109/ROPEC.2016.7830633

Graff M, Miranda-Jimenez S, Tellez ES et al (2020) Evomsa: a multilingual evolutionary approach for sentiment analysis [application notes]. IEEE Comput Intell Mag 15(1):76–88. https://doi.org/10.1109/MCI.2019.2954668

Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional lstm networks. In: Proceedings. 2005 IEEE international joint conference on neural networks, 2005., IEEE, pp 2047–2052. https://doi.org/10.1109/IJCNN.2005.1556215

Grover K, Goel T (2021) Haha@ iberlef2021: Humor analysis using ensembles of simple transformers. In: Proceedings of the Iberian languages evaluation forum (IberLEF 2021), CEUR Workshop Proceedings, Málaga, Spain

Guan Z, Zhou XZ (2021) Tsia at semeval-2021 task 7: Detecting and rating humor and offense. In: Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021), pp 1108–1113. https://doi.org/10.18653/v1/2021.semeval-1.154

Han B, Cook P, Baldwin T (2012) Automatically constructing a normalisation dictionary for microblogs. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, pp 421–432. https://aclanthology.org/D12-1039

Hasan MK, Lee S, Rahman W, et al (2021) Humor knowledge enriched transformer for understanding multimodal humor. In: Proceedings of the AAAI conference on artificial intelligence, pp 12972–12980

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoicka E (2014) The pragmatic development of humor. Prag Develop First Lang Acquisit 10:219

Howard J, Gugger S (2020) Fastai: a layered api for deep learning. Information 11(2):108. https://doi.org/10.3390/info11020108

Howard J, Ruder S (2018) Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146https://doi.org/10.48550/arXiv.1801.06146

Huang HH, Chen CC, Chen HH (2018) Disambiguating false-alarm hashtag usages in tweets for irony detection. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: Short Papers), pp 771–777. https://doi.org/10.18653/v1/p18-2122

Hulstijn J (1996) Automatic interpretation and generation of verbal humor. In: Proceedings o the IWCH96

Ilić S, Marrese-Taylor E, Balazs JA, et al (2018) Deep contextualized word representations for detecting sarcasm and irony. arXiv preprint arXiv:1809.09795https://doi.org/10.48550/arXiv.1809.09795

Ismailov A (2019) Humor analysis based on human annotation challenge at iberlef 2019: first-place solution. In: IberLEF@ SEPLN, pp 160–164

Joshi A, Sharma V, Bhattacharyya P (2015) Harnessing context incongruity for sarcasm detection. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 2: Short Papers), pp 757–762. https://doi.org/10.3115/v1/p15-2124

Jr RWG, O'Brien JE, Doolittle S (1995) Inferring meanings that are not intended: Speakers' intentions and irony comprehension. Discourse Process 20(2):187–203. https://doi.org/10.1080/01638539509544937

Khandelwal A, Swami S, Akhtar SS, et al (2018) Humor detection in english-hindi code-mixed social media content: Corpus and baseline system. arXiv preprint arXiv:1806.05513https://doi.org/10.48550/arXiv.1806.05513

Kui Y (2021) Applying pre-trained model and fine-tune to conduct humor analysis on spanish tweets. In: Proceedings of the Iberian languages evaluation forum (IberLEF 2021), CEUR workshop proceedings, Málaga, Spain

Lan Z, Chen M, Goodman S, et al (2019) Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942https://doi.org/10.48550/arXiv.1909.11942

Liebrecht C, Kunneman F, van den Bosch A (2013) The perfect solution for detecting sarcasm in tweets# not. In: Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis, pp 29–37

Liu F, Weng F, Jiang X (2012) A broad-coverage normalization system for social media language. In: Proceedings of the 50th annual meeting of the association for computational linguistics (volume 1: Long Papers), pp 1035–1044. https://aclanthology.org/P12-1109

Liu P, Li W, Zou L (2019a) Nuli at semeval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In: SemEval@ NAACL-HLT, pp 87–91, https://doi.org/10.18653/v1/s19-2011

Liu Y, Ott M, Goyal N, et al (2019b) Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692https://doi.org/10.48550/arXiv.1907.11692

Loper E, Bird S (2002) Nltk: the natural language toolkit. arXiv preprint cs/0205028 https://doi.org/10.48550/arXiv.cs/0205028

Lozano-Palacio I, de Mendoza Ibáñez FJR (2022) Modeling Irony: a cognitive-pragmatic account. John Benjamins

Mao J, Liu W (2019) A bert-based approach for automatic humor detection and scoring. In: IberLEF@ SEPLN, pp 197–202

Maynard DG, Greenwood MA (2014) Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In: Lrec 2014 proceedings, ELRA

Meaney J, Wilson SR, Chiruzzo L, et al (2021) Semeval 2021 task 7, hahackathon, detecting and rating humor and offense. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing. https://doi.org/10.18653/v1/2021.semeval-1.9

Meyer JC (2000) Humor as a double-edged sword: four functions of humor in communication. Commun Theory 10(3):310–331. https://doi.org/10.1111/j.1468-2885.2000.tb00194.x

Miller T, Do Dinh EL, Simpson E, et al (2019) Ofai-ukp at haha@ iberlef2019: predicting the humorousness of tweets using gaussian process preference learning. In: IberLEF@ SEPLN, pp 180–190

Miyato T, Dai AM, Goodfellow I (2016) Adversarial training methods for semi-supervised text classification. arXiv preprint arXiv:1605.07725https://doi.org/10.48550/arXiv.1605.07725

Mondal A, Sharma R (2021) Team_kgp at semeval-2021 task 7: a deep neural system to detect humor and offense with their ratings in the text data. In: Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021), pp 1169–1174. https://doi.org/10.18653/v1/2021.semeval-1.164

Nanda A, Singh AP, Gupta A, et al (2021) Techssn at haha@ iberlef 2021: Humor detection and funniness score prediction using deep learning. In: Proceedings of the Iberian languages evaluation forum (IberLEF 2021), CEUR workshop proceedings, Málaga, Spain

Ortega-Bueno R, Muniz-Cuza CE, Pagola JEM, et al (2018) Uo upv: Deep linguistic humor detection in spanish social media. In: Proceedings of the third workshop on evaluation of human language technologies for Iberian languages (IberEval 2018) co-located with 34th conference of the Spanish society for natural language processing (SEPLN 2018), pp 204–213

Ortega-Bueno R, Rosso P, Pagola JEM (2019) Uo upv2 at haha 2019: Bigru neural network informed with linguistic features for humor recognition. In: Proceedings of the Iberian languages evaluation forum (IberLEF 2019). CEUR workshop proceedings, CEUR-WS, Bilbao, Spain (9 2019)

Ortiz-Bejar J, Salgado V, Graff M, et al (2018) Ingeotec at ibereval 2018 task haha: μtc and evomsa to detect and score humor in texts. In: Proceedings of the third workshop on evaluation of human language technologies for Iberian languages (IberEval 2018) co-located with 34th conference of the Spanish sSociety for natural language processing (SEPLN 2018)

Ortiz-Bejar J, Tellez ES, Graff M, et al (2019) Ingeotec at iberlef 2019 task haha. In: IberLEF@ SEPLN, pp 203–211

Padró L, Stanilovsky E (2012) Freeling 3.0: towards wider multilinguality. In: LREC2012. http://hdl.handle.net/2117/15986

Pamungkas EW, Patti V (2018) # nondicevosulserio at semeval-2018 task 3: Exploiting emojis and affective content for irony detection in english tweets. In: International workshop on semantic evaluation, Association for Computational Linguistics, pp 649–654. https://doi.org/10.18653/v1/s18-1106

Pannu A (2015) Artificial intelligence and its application in different areas. Artif Intell 4(10):79–84

Peña MS, Ruiz de Mendoza FJ (2017) Construing and constructing hyperbole. Stud Figurative Thought Lang 56:41

Pennebaker JW, Francis ME, Booth RJ (2001) Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates 71(2001):2001

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. https://doi.org/10.3115/v1/d14-1162

Peters ME, Neumann M, Iyyer M, et al (2018) Deep contextualized word representations. In: Proceedings of the NAACL. https://doi.org/10.18653/v1/n18-1202

Raghunathan A, Xie SM, Yang F, et al (2019) Adversarial training can hurt generalization. arXiv preprint arXiv:1906.06032https://doi.org/10.48550/arXiv.1906.06032

Rangwani H, Kulshreshtha D, Singh AK (2018) Nlprl-iitbhu at semeval-2018 task 3: combining linguistic features and emoji pre-trained cnn for irony detection in tweets. In: Proceedings of the 12th international workshop on semantic evaluation, pp 638–642. https://doi.org/10.18653/v1/s18-1104

Reyes A, Rosso P, Buscaldi D (2012) From humor recognition to irony detection: The figurative language of social media. Data Knowl Eng 74:1–12. https://doi.org/10.1016/j.datak.2012.02.005, www.sciencedirect.com/science/article/pii/S0169023X12000237, applications of Natural Language to Information Systems

Reyes A, Rosso P, Buscaldi D (2012) From humor recognition to irony detection: the figurative language of social media. Data Knowl Eng 74:1–12. https://doi.org/10.1016/j.datak.2012.02.005

Reyes A, Rosso P, Veale T (2013) A multidimensional approach for detecting irony in twitter. Lang Res Eval 47(1):239–268. https://doi.org/10.1007/s10579-012-9196-x

Ribeiro MT, Singh S, Guestrin C (2016) " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 1135–1144. https://doi.org/10.1145/2939672.2939778

Rohanian O, Taslimipoor S, Evans R, et al (2018) Wlv at semeval-2018 task 3: Dissecting tweets in search of irony. Association for Computational Linguistics. https://doi.org/10.18653/v1/s18-1090

Ruiz de Mendoza F, Lozano I (2021) On verbal and situational irony: towards a unified approach. https://doi.org/10.1075/ftl.11.07rui

Sennrich R, Haddow B, Birch A (2015) Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909https://doi.org/10.48550/arXiv.1508.07909

Sharma C, Bhageria D, Scott W, et al (2020) Semeval-2020 task 8: Memotion analysis–the visuo-lingual metaphor! arXiv preprint arXiv:2008.03781https://doi.org/10.48550/arXiv.2008.03781

Song B, Pan C, Wang S, et al (2021) Deepblueai at semeval-2021 task 7: Detecting and rating humor and offense with stacking diverse language model-based methods. In: Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021), pp 1130–1134. https://doi.org/10.18653/v1/2021.semeval-1.158

Sperber D, Wilson D (1981) Irony and the use-mention distinction. Philosophy 3:143–184

Swamy SD, Jamatia A, Gambäck B, et al (2019) Nit_agartala_nlp_team at semeval-2019 task 6: An ensemble approach to identifying and categorizing offensive language in twitter social media corpora. In: NAACL HLT 2019 the international workshop on semantic evaluation proceedings of the thirteenth workshop, Association for Computational Linguistics. https://doi.org/10.18653/v1/s19-2124

Swamy SD, Laddha S, Abdussalam B, et al (2020) Nit-agartala-nlp-team at semeval-2020 task 8: building multimodal classifiers to tackle internet humor. arXiv preprint arXiv:2005.06943https://doi.org/10.48550/arXiv.2005.06943

Tasneem F, Naim J, Chy AN (2020) Harnessing ensemble of data preprocessing and hand-crafted features for irony detection in tweets. In: 2020 23rd international conference on computer and information technology (ICCIT), IEEE, pp 1–6, https://doi.org/10.1109/iccit51783.2020.9392711

Tay Y, Zhang A, Tuan LA, et al (2019) Lightweight and efficient neural natural language processing with quaternion networks. arXiv

preprint arXiv:1906.04393https://doi.org/10.48550/arXiv.1906.04393
Tay Y, Dehghani M, Bahri D, et al (2020) Efficient transformers: a survey. arXiv preprint arXiv:2009.06732https://doi.org/10.48550/arXiv.2009.06732
Tellez ES, Miranda-Jiménez S, Graff M, et al (2017) A simple approach to multilingual polarity classification in twitter. Pattern Recogn Lett 94:68–74. https://doi.org/10.48550/arXiv.1612.05270
Tellez ES, Moctezuma D, Miranda-Jiménez S, et al (2018) An automated text categorization framework based on hyperparameter optimization. Knowl Based Syst 149:110–123. https://doi.org/10.48550/arXiv.1704.01975
Tobin V, Israel M (2012) Irony as a viewpoint phenomenon. In: Viewpoint in language: a multimodal perspective, pp 25–46
Tomás D, Ortega-Bueno R, Zhang G, et al (2022) Transformer-based models for multimodal irony detection. J Ambient Intell Human Comput, pp 1–12
Tsipras D, Santurkar S, Engstrom L, et al (2018) Robustness may be at odds with accuracy. arXiv preprint arXiv:1805.12152https://doi.org/10.48550/arXiv.1805.12152
van den Beukel S, Aroyo L (2018) Homonym detection for humor recognition in short text. In: Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis, pp 286–291. https://doi.org/10.18653/v1/w18-6242
Van Hee C, Lefever E, Hoste V (2018) Semeval-2018 task 3: Irony detection in english tweets. In: Proceedings of The 12th international workshop on semantic evaluation, pp 39–50. https://doi.org/10.18653/v1/s18-1005
Wang SI, Manning CD (2012) Baselines and bigrams: Simple, good sentiment and topic classification. In: Proceedings of the 50th annual meeting of the association for computational linguistics (volume 2: Short Papers), pp 90–94. https://aclanthology.org/P12-2018
Weller O, Seppi K (2019) Humor detection: a transformer gets the last laugh. arXiv preprint arXiv:1909.00252https://doi.org/10.48550/arXiv.1909.00252
Wilson D (2006) The pragmatics of verbal irony: Echo or pretence? Lingua 116(10):1722–1743. https://doi.org/10.1016/j.lingua.2006.05.001. (Language in mind: a tribute to Neil Smith on the Occasion of his Retirement)
Wilson D, Sperber D (1992) On verbal irony. Lingua 87(1):53–76
Wilson D, Sperber D (2012) Explaining irony. Meaning Relevance, pp 123–145
Wu C, Wu F, Wu S, et al (2018) Thu_ngn at semeval-2018 task 3: Tweet irony detection with densely connected lstm and multi-task learning. In: Proceedings of The 12th international workshop on semantic evaluation, pp 51–56. https://doi.org/10.18653/v1/s18-1006
Zampieri M, Malmasi S, Nakov P, et al (2019) Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). arXiv preprint arXiv:1903.08983https://doi.org/10.48550/arXiv.1903.08983
Zhang M, Yang J, Teng Z, et al (2016) Libn3l: a lightweight package for neural nlp. In: Proceedings of the tenth international conference on language resources and evaluation (LREC'16), pp 225–229. https://aclanthology.org/L16-1034
Zhao X, Xu B, Zheng D, et al (2018) Tweet irony detection using ensembles of word level attentive long short-term memory and convolutional neural network. In: 2018 14th international conference on natural computation, fuzzy systems and knowledge discovery (ICNC-FSKD), IEEE, pp 524–529. https://doi.org/10.1109/fskd.2018.8687128
Zhao Y, Tao X (2021) Zyj at semeval-2021 task 7: Hahackathon: Detecting and rating humor and offense with albert-based model. In: Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021), pp 1175–1178. https://doi.org/10.18653/v1/2021.semeval-1.165
Zhou C, Sun C, Liu Z, et al (2015) A c-lstm neural network for text classification. arXiv preprint arXiv:1511.08630https://doi.org/10.48550/arXiv.1511.08630
Zylich B, Gugnani A, Brookman G, et al (2021) Amherst685 at semeval-2021 task 7: joint modeling of classification and regression for humor and offense. In: Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021), pp 1190–1195. https://doi.org/10.18653/v1/2021.semeval-1.168