Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 73-83 recibido 14-12-2022 revisado 05-02-2023 aceptado 13-02-2023

On the Poor Robustness of Transformer Models in Cross-Language Humor Recognition

Sobre la Poca Robustez de los Modelos Transformers en el Reconocimiento Translingüístico del Humor

Roberto Labadie Tamayo[1] **Reynier Ortega-Bueno**[1]
Paolo Rosso[1] **Mariano Rodriguez Cisneros**[2]
[1]Universitat Politècnica de València
[2]Harbour.Space University, Barcelona
rlabtam@posgrado.upv.es, rortega@prhlt.upv.es, prosso@dsic.upv.es,
mjasonrc@gmail.com

Abstract: Humor is a pervasive communicative device; nevertheless, its portability from one language to another remains challenging for computer machines and even humans. In this work, we investigate the problem of humor recognition from a cross-language and cross-domain perspective, focusing on English and Spanish languages. To this aim, we rely on two strategies: the first is based on multilingual transformer models for exploiting the cross-language knowledge distilled by them, and the second introduces machine translation to learn and make predictions in a single language. Experiments showed that models struggle in front of the humor complexity when it is translated, effectively tracking a degradation in the humor perception when messages flow from one language to another. However, when multilingual models face a cross-language scenario, exclusive between the fine-tuning and evaluation data languages, humor translation helps to align the knowledge learned in fine-tuning phase. According to this, a mean increase of 11% in F1 score was observed when classifying English-written texts with models fine-tuned with a Spanish dataset. These results are encouraging and constitute the first step towards a computationally cross-language analysis of humor.
Keywords: humor recognition, humor translation, cross-language humor, multilingual models.

Resumen: El humor es un recurso comunicativo muy extendido; sin embargo, su portabilidad de un idioma a otro sigue siendo un reto para las máquinas informáticas e incluso para los humanos. En este trabajo, investigamos el problema del reconocimiento del humor desde una perspectiva translingüística y transdominio. Para ello, recurrimos a dos estrategias: la primera se basa en modelos transformers multilingües para explotar el conocimiento translingüístico que son capaces de destilar, y la segunda introduce la traducción automática para aprender y hacer predicciones en un solo idioma. Los experimentos demostraron que los modelos tienen dificultades ante la complejidad del humor cuando se traduce, lo que supone una degradación de la percepción del humor cuando los mensajes pasan de un idioma a otro. Sin embargo, cuando los modelos multilingües se enfrentan a un escenario translingüístico, exclusivo entre los idiomas de los datos de refinado y de evaluación, la traducción del humor ayuda a alinear los conocimientos aprendidos en la fase de refinado. En consecuencia, se observó un aumento medio del 11% de la puntuación F1 al clasificar textos escritos en inglés con modelos refinados con un conjunto de datos en español. Estos resultados son alentadores y constituyen el primer paso hacia un análisis computacional multilingüe del humor.
Palabras clave: detección de humor, traducción del humor, humor translingüe, modelos multilingües.

ISSN 1135-5948 DOI 10.26342/2023-70-6

1 Introduction

There is a set of evolved emotional functions shared by humans; laughter is part of this universal language of basic emotions that all humans recognize (Savage et al., 2017). Nevertheless, despite its ubiquity, proper comprehension of some humorous expressions goes beyond the semantics involved in messages. It relies on information from the context where jokes are made and the receptor's background knowledge (Tsakona, 2017), which implies a different or even null perception from one person to another. Moreover, when it comes to such creative device as humor, language plays a critical role in perceiving the funny meaning. Particularly, when information flows from one language to another on its way to the receptor, a joke's intended meaning is at risk of vanishing.

Wordplays are examples of language-dependent expressions that can be potentially misunderstood upon literal translation into a different language since they employ the arrangement and phonetics of words to produce humor. For example, in:
Why do male ants float while female ants sink? They're buoy-ant
It is very challenging to translate the phrase to ensure humor understanding by a non-English speaker, regardless of their background knowledge. Whereas, in the case of:

A: *Are you already here?*

B: *No, I'm just a figment of your imagination.*

The literal translation can still provoke laughter.

On the other hand, linguistic diversity on the Internet increases due to its interconnecting nature (Paolillo, 2007). In social media, where people from different cultural backgrounds and ethnicities share information, dealing with this multilingual phenomenon is inherent when identifying and filtering content and behaviors appropriated for specific users.

Many Natural Language Processing (NLP) tasks have been covered from a multilingual perspective in the scenario of social media with machine learning models (Ghanem et al., 2020; Wang et al., 2019; Al-Hassan and Al-Dossari, 2019). Most works tackle the under-representation of some languages by extending the knowledge learned from one language to another. In this sense, multilingual transformer-based architectures have become the state of the art in almost all of them (Wang et al., 2020; Chauhan et al., 2022). Despite the growing interest in humor in many languages such as English (Ermakova et al., 2022a; Meaney et al., 2021; Hossain et al., 2020), Spanish (Chiruzzo et al., 2021), Portuguese (Clemêncio, Alves, and Gonçalo Oliveira, 2019), and Chinese (Wu et al., 2021), to the best of our knowledge few efforts have been made to investigate the task of humor recognition from a computational cross-domain and cross-language perspective.

Machine translation paves the way for facing the challenge of multilingualism in texts[1]. Although these tools have been adequate for translating literal texts, when dealing with figurative language their performance drop considerably. In fact, humorous texts that often appeal to cultural knowledge or play on words become a complex problem in Machine Translation (Attardo, 2002; Zabalbeascoa, 2005; Popa, 2005; Low, 2011). Despite those shortcomings, we consider that some types of *self-contained* funny texts could preserve their meaning from one language to another. That is, the humor purely related to semantics without requiring additional cultural knowledge or information from the context. Moreover, we think that some linguistic features, not necessarily associated with the semantics and pragmatics involved in texts, may help to recognize humor without the need of understanding the text's whole meaning.

In light of the facts above, we consider that more efforts must be paid to investigate humor recognition in cross-domain and cross-languages scenarios. Particularly, we aim at stressing both Machine Translation systems and multilingual transformer-based models in order to identify their feasibility in humor recognition across languages. For that, we address the following research questions:

RQ1. What is the impact of machine translation on the semantics of humorous messages?

RQ2. How robust are multilingual trans-

[1]https://syncedreview.com/2020/05/20/neural-network-ai-is-the-future-of-the-translation-industry/

former models when dealing with translated humorous messages?

RQ3. Is it better to work with the same language of the dataset employed to fine-tune the multilingual transformer model for recognising humor (by automatically translating the target language)?

In RQ1 we aim at investigating how the semantics of humorous messages change upon automatic translation and how transformer models perceive this change. For RQ2, we will study the capability of multilingual transformer models to recognize specifically the presence of humor in translated messages. In RQ3, we are interested in investigating if in the case of multilingual transformers fine-tuned to recognize humor with English-written messages and evaluated on Spanish samples, it is better to automatically translate these samples into English. The latter, also for the opposite direction, when using multilingual transformers fine-tuned with Spanish-written messages and evaluated on English samples.

This paper presents a study on the behavior of transformer-based neural models for addressing humor recognition from a cross-language perspective. We take into account the *self-contained* and *language-dependent* humor phenomena, e.g. puns. Regarding the latter, we explore if the self-contained humor recognition methodology is extensible to its language-dependent kind. The rest of the paper is organized as follows: Section 2 presents some related works on humor recognition also from multilingual and cross-language perspectives. In Section 3, we describe the data and strategies studied as well as the employed methodology. In Section 4, we describe the experimental setup and results. Finally, in Section 5, we discuss the results achieved and provide some directions to explore as further work.

2 Related Works

Computational humor recognition is a widely explored issue. One of the first empirical pieces of evidence of this task's feasibility were given by (Mihalcea and Strapparava, 2005). From there on, several works have been conducted to integrate contextual, visual, and acoustic information in multimodal approaches (Yang, Ai, and Hirschberg, 2019; Vásquez and Aslan, 2021; Song et al., 2021; Chauhan et al., 2022; Tomás et al., 2022). Nevertheless, just a few works examine the phenomenon of humor from a cross-language and multilingual view (Chauhan et al., 2022).

Systems based on large pre-trained language models have outperformed the state of the art in many NLP tasks, including humor recognition (Grover and Goel, 2021; Subies, Sánchez, and Vaca, 2021) and machine translation (Vaswani et al., 2017).
However, humor translation remains a field with a huge room for improvement due to its subjectivity and linguistic complexity. Some of the most recent works (Miller, 2019) provide an interactive method for the computer-assisted translation of puns.

In this line, the task JOKER@CLEF 2022: Automatic Wordplay and Humour Translation Workshop (Ermakova et al., 2022b), where participants were asked to perform translation of humorous texts and identify its nature, was the first attempt to construct a parallel and multilingual humor corpus. Here, most participants' approaches relied again on transformers-based models, this time reinforced with templates featuring (Arroubat, 2022; Anne-Gwenn et al., 2022).

Besides the poor existence of parallel corpora, the above-referred issues in humor translation and recognition have worsened the scarcity of work transferring humor knowledge from one language to another.

3 Methodology

Evaluating the robustness of transformer models from a cross-language perspective for our task requires a parallel corpus with humorous information. Nevertheless, its absence forced us to compile a corpus considering different sources and languages. For this purpose, we focus on the English and Spanish languages, given the extensive amount of work related to humor recognition on them.

Taking into account the growing application of pre-trained transformers models in almost every NLP task, we employ three multilingual variants to evaluate their performance. However, as this work is not intended to outperform the SOTA in humor recognition, we simply stack a ReLU-activated layer between the encoder module and a softmax classification layer for each model. Then, the models are trained and evaluated with

an end-to-end fashion by feeding the ReLU-activated layer with the [CLS] vector from the last encoder block of the transformer (more details in Section 4). We employ multilingual models since we hypothesize they allow to capture and share background knowledge regardless of the language used during the fine-tuning process and the evaluation.

The first model, (*ml-base*) BERT-multilingual-base (Devlin et al., 2018), was pre-trained on the top 104 languages with the largest Wikipedia using a masked language modeling (MLM) and next sentence prediction (NSP) objectives. The second, (*ml-sentiment*), is a fine-tuned version of the latter in a sentiment analysis task on texts from six languages[2], among them English and Spanish, which are the ones addressed in this work. We use this model because although its pre-training knowledge comes from (*ml-base*), the information introduced by the sentiment-tuning could provide us with criteria diversity to characterize the general behavior of humor empirically. Finally, we study another variation of the BERT-base (*ml-distil*) model into a smaller and distilled architecture proposed by (Sanh et al., 2019), trained on the top 104 languages with the largest Wikipedia.

The source code and datasets employed in this study are publicly available in GitHub[3] for reproducibility.

3.1 Datasets

We gathered the monolingual datasets of 4 shared tasks:

(i) SemEval-2020 Task 7: Assessing Humor in Edited News Headlines (Hossain et al., 2020).

(ii) SemEval 2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense (Meaney et al., 2021).

(iii) HAHA at IberLEF 2021: Detecting, Rating and Analyzing Humor in Spanish (Chiruzzo et al., 2021).

(iv) JOKER@CLEF 2022 Task 1: Classify and Explain Instances of Wordplay (Ermakova et al., 2022a).

These datasets are annotated with several aspects related to humor, including whether it is present or not. However, they assess it with texts of different genres and writing styles, including tweets, headlines, or isolated wordplays, and representing different knowledge domains. This enables us to see two perspectives of the aggregation: the language-level, where datasets in the same language are grouped into a single corpus, and the domain-level, where each dataset, regardless of its language, is analyzed separately.

SemEval-2020 Task 7 Dataset

For this task, given a headline in the English language and a micro-edited version (*viz.*, replacement of entity by noun, noun by noun, verb by verb), participants were asked to determine whether this substitution generates a funny message. In the dataset (*Headlines*), for each headline, the replacement was annotated as well as the humor rating given by 6 annotators on a 0 to 3 scale and its mean value. From here, we consider as negative examples the original headline and as positive examples of humor those micro-editions whose mean humor rating was above 2 (*i.e.,* moderately funny and funny).

SemEval 2021 Task 7 Dataset

The dataset from this task (*Hahackathon*) contains English texts from Twitter and the Kaggle Short Jokes dataset, described with the presence of humor as well as humor rating, controversiality, and offensiveness rating of the messages by 20 different annotators. In this work, we only focus on the binary annotation regarding whether a text can be considered as funny.

IberLEF 2021 HAHA Dataset

In the shared task HAHA 2021, it was proposed a dataset (*HAHA*) composed of tweets written in Spanish, annotated regarding the presence of humor, funniness score, the humor mechanism employed (*e.g.,* parody, stereotype, etc.), and the humor target, *i.e.,* for a humorous tweet, the target of the joke from a set of classes such as racist jokes, sexist jokes, etc. We are interested in the binary annotation of humor, even when the remaining annotation is valuable for refining the humor analysis considering the language mechanism, the purpose, and the victims of

[2]https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

[3]https://github.com/labadier/Humor.git

the jokes.

JOKER@CLEF 2022 Task 1 Dataset
For this task, given a wordplay in the English language, participants were asked to classify it, attending different criteria. They also must identify and disambiguate the target words as an explanation of the wordplay. In the dataset (*JOKER*), the criteria annotated for each example include whether the source and the target of the wordplay co-occur in the text (horizontal/vertical), the manipulation type, *viz.* identity, similarity, permutation, abbreviation, if cultural reference is needed in order to understand the instance of wordplay, whether it is offensive or not, and whether the wordplay is in conventional form. Also, the target words and the disambiguation of the wordplay are annotated.

From these data examples, we just take those whose manipulation is by permutation (the textual material is given a new order, as in anagrams or spoonerism. e.g. "Dormitory = dirty room"), similarity (source and target are not perfectly identical, but the resemblance is obvious, e.g. "They're called lessons because they lessen from day to day") or Identity (source and target are formally identical, e.g. "How do you make a cat drink? Easy: put it in a liquidizer").

Table 1 shows the distribution of the examples in each dataset. The balance between positive (humor) and negative (non-humor) classes follows the one proposed by their authors. For training and testing the models from a cross-language perspective, we assume two partitions, one composed of examples originally in English from *SemEval 2021 Task 7 Dataset*, *JOKER@CLEF 2022 Task 1 Dataset* and *SemEval-2020 Task 7 Dataset*, with a representation of the positive class at training of 43%. The second is represented just by *IberLEF 2021 HAHA Dataset* dataset with a 39% of humorous examples for training. The instances in both partitions come from different knowledge domains and humor styles; then, besides the cross-language difficulty, we also have to deal with cross-domain data during evaluation.

To answer the first question, we first study how the humorous perception of these basic models varies for a back-translated instance. Thus, we check if the semantic underlying in the texts (Attardo, 2017) is preserved, even when their humorous incongruity vanishes in the pivot language during back-translation.

We rely on this experiment to disaggregate two sources of errors in the prediction stage when studying RQ2. The first is the one related to the learned parameters and model's architecture for recognizing humor. The second comes when instances are translated to the same language of the samples the multilingual transformer has been fine-tuned with.

Once we determine the impact of the noise introduced directly by machine translation, we can dive into RQ2. For this, we evaluate how translating texts impacts humor recognition in a cross-language scenario under multilingual models.

For evaluating each strategy, we use Micro-F1 over the positive class taking into account that at the domain-level, there are cases of slight unbalance or extreme scenarios where there are no examples for the non-humor class, as in *JOKER*.

4 Experimental Results

In the fine-tuning process of every model, we optimized the parameters with the RMSprop algorithm (Hinton, Srivastava, and Swersky, 2012) by employing an increasing learning rate from the shallower layers to the deeper ones (Howard and Ruder, 2018), starting from 1e-5 and increasing it on each layer with a factor of 0.1 units.

Every translation step involved in this work was accomplished with googletrans library, using Spanish as the complementary language for English, and vice versa. In the same way, when we study approaches based

Language	Dataset	Train ($\mathcal{T}$)		Test ($\mathcal{D}$)	
		Humor	Non-Humor	Humor	Non-Humor
English	*SemEval 2021 Task 7*	3436	5564	385	615
	JOKER@CLEF 2022 Task 1	531	0	4516	0
	SemEval-2020 Task 7 Dataset	890	890	88	88
Spanish	*IberLEF 2021 HAHA*	11595	18405	3000	3000

Table 1: Statistics of the datasets.

on back-translation, this complementarity relation is applied to select the pivot language.

As we mentioned before, the first step is to study how noisy machine translation results for humor regarding the semantics of the message; for this, we applied back-translation over the instances and investigated how humor perception vanishes for every multilingual model. In Table 2 are shown the results in terms of F1 for every model (detailed at *domain-level*), where $\mathcal{D}$ stands for the original version of our test sets described in Table 1, $\mathcal{D}^*$ is the version of $\mathcal{D}$ where each instance is translated into its complementary language and $\mathcal{D}^{**}$ corresponds to the back-translated version of $\mathcal{D}$. We include an estimation of the F1 95%-confidence interval (ci) by Percentile bootstrapping according to (DiCiccio and Efron, 1996).

Model	Dataset	$\mathcal{D}$	ci	$\mathcal{D}^{**}$
ml-base	*Hahack.*	0.921	0.015	0.923
	JOKER	0.941	0.005	0.939
	Headlines	0.778	0.062	0.772
	HAHA	0.870	0.008	0.869
ml-sent	*Hahack.*	0.914	0.017	0.916
	JOKER	0.934	0.005	0.933
	Headlines	0.814	0.050	0.802
	HAHA	0.871	0.008	0.870
ml-distil	*Hahack.*	0.905	0.018	0.903
	JOKER	0.945	0.005	0.944
	Headlines	0.709	0.070	0.716
	HAHA	0.863	0.009	0.861

Table 2: Variation in humor perception by multilingual transformer models after back-translation.

Here we can see that the error in $\mathcal{D}^{**}$ (a perturbed instance of $\mathcal{D}$) is not statistically significant w.r.t. the results on $\mathcal{D}$ if we assume the learned parameters of the models.

Since we only seek an empirical probe of the model's capability to find a similar interpretation of the back-translated data w.r.t. the original, for this experiment we train every multilingual model by employing all the domains and languages at the same time, allowing the knowledge-sharing among all the *domain-level* datasets.

However, we explored how this knowledge-sharing impacts the results for the cross-domain scenario present in the English *language-level* dataset. Table 3 shows different domain combinations for fine-tuning *ml-base* model, where K, H and J, refers to *Hahackathon*, *Headlines* and *JOKER* respectively[4].

Setting	Test Set		
	JOKER	Headlines	*Hahack.*
H	-	0.737	-
K	-	-	0.913
K+H	0.713	-	-
K+J	-	0.667	-
H+J	-	-	0.764
K+H+J	0.906	0.749	0.920

Table 3: Cross-domain settings for English datasets.

As we can see, using a purely cross-domain scenario (rows 3-5) has a negative impact on the model's performance. Nevertheless, when this external knowledge is used as a way of data augmentation (last row), it effectively helped to improve the achieved results. We can notice that in all cases, the results are inferior with respect to those obtained in Table 2, even when the fine-tuning is carried out across all the domains in the English language (last row). The latter suggests that the model employs knowledge from *HAHA* (Spanish corpus) to make inferences in English-written texts. Considering that, we investigated the effectiveness of using a multilingual system in a cross-language scenario by means of a zero-shot approach. We fine-tuned every model with the data from the English *language-level* dataset to evaluate the data from the Spanish *language-level* dataset and vice versa. Table 4 shows the results obtained in each case.

If we compare the results from Table 4 obtained using *ml-bert* with those from Table 3, we can observe that the model performance diminishes in each dataset. This suggests a greater contribution from the cross-domain knowledge for humor recognition with the studied transformer-based models.

Once we have studied the cross-language and cross-domain impact on humor recognition, we can explore how feasible it is to extend the knowledge by means of translation at the evaluation phase. Also, given the

[4]We were not able to evaluate the model trained on the JOKER dataset since it only consists of positive examples of humor.

Fine-tuning Language	Dataset	ml bert	ml sentiment	ml distil
Spanish	*Hahackathon*	0.760	0.753	0.754
	JOKER	0.666	0.661	0.650
	Headlines	0.534	0.528	0.500
English	*HAHA*	0.754	0.713	0.729

Table 4: Cross-language scenario results.

results of Table 2, where we found machine translation did not distort the semantics of funny texts, we are able to explore the issues of the humor recognition systems regardless of any possible *meaning changes* introduced by machine translation.

4.1 Humor Recognition in Translated Instances

As described in the strategy for study RQ3 in Section 1, we introduce a cross-language scenario again, but this time we tried to mitigate it by translating the evaluation instances into the fine-tuning language. Table 5 shows the results of the evaluation in these translated $\mathcal{D}^*$ datasets.

Dataset ($\mathcal{D}^*$)	ml bert	ml sentiment	ml distil
Hahackathon	0.808	0.825	0.787
JOKER	0.736	0.719	0.743
Headlines	0.553	0.554	0.512
HAHA	0.767	0.734	0.731

Table 5: Language inversion to reduce cross-language effect.

Here, we can see an improvement with respect to the previous results in Table 4, which means at least some of the humorous perception is preserved after translation and makes more useful the information learned during the model fine-tuning process.

The latter studies do not allow us to isolate the vanishing of humor recognition introduced when instances are translated. To this end, for evaluating the $\mathcal{D}^*$ dataset, we employed the same model parameters from the experiments referred to in Table 2, where cross-domain knowledge sharing was allowed.

Looking over the results from Table 6 with respect to $\mathcal{D}$ and $\mathcal{D}^{**}$ in Table 2, we can observe a poor robustness of the transformer models associated to humor translation. In

Model	Dataset	$\mathcal{D}^*$
ml-base	*Hahack.*	0.880
	JOKER	0.875
	Headlines	0.659
	HAHA	0.811
ml-sent	*Hahack.*	0.856
	JOKER	0.861
	Headlines	0.641
	HAHA	0.803
ml-distil	*Hahack.*	0.833
	JOKER	0.885
	Headlines	0.616
	HAHA	0.789

Table 6: Results for evaluation in translated instances.

the prediction phase, the models had in common issues associated with polysemous words, phrase ambiguities from the source language as regards the target language, and word rearrangements, particularly in wordplays. Table 7 shows examples from *Hahackathon* and *HAHA* related to this problem.

India is a very peaceful country because nobody has any beef over there. India es un país muy pacífico porque nadie tiene problemas allí.
Two dyslexics walk into a bra Dos disléxicos entran en un sostén
—¿Follamos? —No, que yo recuerde. —"Shall we fuck?" -Not that I remember.

Table 7: Translation ambiguities examples.

In the case of the *Headlines* dataset, which exhibits the greater drop in performance, it can be noticed that besides

the translation degeneration, examples are culturally dependent and related to knowledge and vocabulary distant from the one employed in the pre-training and fine-tuning phase of the evaluated models [5]. That is, HAHA vocabulary represents informal Twitter texts, and *Headlines* involves in some way "journalistic" and more formal vocabulary. Table 8 shows some examples related to the *Headlines* phenomenon.

Gov. Kasich slams President Trump's move on haircut care subsidies
White House spokesman does not rule out Trump-Putin July cuddling in Germany

Table 8: Contextual Dependency of HAHA translated examples.

Experiments developed in this section showed that humor translation helps the model to extend the knowledge learned in one language for inference in examples written in another one, i.e., it helps to mitigate the cross-language effect in some cases. Nevertheless, these models still struggle in front of the humor complexity as a communicative device when it is translated, effectively tracking a degeneration in the humor perception when messages flow from one language to another.

5 Conclusions and Future Work

Humor relies on the incongruences of two semantic planes that, when contrasted by the receptor, produce it in a natural way. Its translation comes with different implications that make pre-trained transformer-based models not robust to recognize it in a cross-language scenario. The main concerns are related to contextual information, background knowledge dependency, and lexical characteristics of the language (RQ2). This vanishing becomes more severe in creative ways of humor, such as wordplays involving phonetics, word polysemy, and phrasal ambiguity. Nevertheless, neural machine translation is capable of individually preserving the humorous semantics, as we examined in our work (RQ1). Also, despite the referred humor recognition vanishing, when we translate and evaluate the samples directly in the language of the models' fine-tuning process, they achieve better performance for recognizing humor in a cross-language scenario (RQ3).

As future work, we plan to extend this analysis towards a broader range of languages and translations provided by ready available machine translation systems to ensure reproducibility. Moreover, since almost every top-ranked system proposed in the shared tasks related to the explored datasets employed transformer-based architectures, we plan to evaluate their proposal on the experiments presented in this study as a way of obtaining more empirical evidence.

Finally, as cultural and contextual knowledge plays an important role in the performance decrease, we plan to explore two strategies. The first is to study how mitigating topic bias in datasets helps the model to address the cross-domain phenomenon. The second strategy consists of partially updating the knowledge of models by determining key examples as domain concepts from the new datasets and incorporating them when fine-tuning the model.

Acknowledgments

This work has been partially developed with the support of valgrAI - Valencian Graduate School and Research Network of Artificial Intelligence and the Generalitat Valenciana, and co-funded by the European Union. The work of Ortega Bueno and Rosso was in the framework of the FairTransNLP research project (PID2021-124361OB-C31) funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe.

References

Al-Hassan, A. and H. Al-Dossari. 2019. Detection of hate speech in social networks: a survey on multilingual corpus. In *6th International Conference on Computer Science and Information Technology*, volume 10, pages 10–5121.

Anne-Gwenn, B., E. Liana, D. de Saint-Cyr Florence, D. L. Pierre, C. Victor, P.-H. Nicolas, A. Benoit, A. Jean-Victor, D. Alexandre, G. Juliette, H. Aymeric, and M.-B. Florian. 2022. Poetic or humorous text generation: Jam event at

[5] In these cases models were fine-tuned with data originally in Spanish (*HAHA*).

pfia2022. In G. Faggioli, N. Ferro, A. Hanbury, and M. Potthast, editors, *Proceedings of the Working Notes of CLEF 2022: Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings.

Arroubat, H. 2022. Wordplay location and interpretation with deep learning methods. In G. Faggioli, N. Ferro, A. Hanbury, and M. Potthast, editors, *Proceedings of the Working Notes of CLEF 2022: Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings.

Attardo, S. 2002. Translation and Humour. *The Translator*, 8(2):173–194.

Attardo, S., 2017. *The General Theory of Verbal Humor*, chapter chapter10. Routledge.

Chauhan, D. S., G. V. Singh, A. Arora, A. Ekbal, and P. Bhattacharyya. 2022. A sentiment and emotion aware multimodal multiparty humor recognition in multilingual conversational setting. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6752–6761, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Chiruzzo, L., S. Castro, S. Góngora, A. Rosa, J. A. Meaney, and R. Mihalcea. 2021. Overview of haha at iberlef 2021: Detecting, rating and analyzing humor in spanish. *Procesamiento del Lenguaje Natural*, 67(0):257–268.

Clemêncio, A., A. Alves, and H. Gonçalo Oliveira. 2019. Recognizing humor in portuguese: First steps. In *Progress in Artificial Intelligence: 19th EPIA Conference on Artificial Intelligence, EPIA 2019, Vila Real, Portugal, September 3–6, 2019, Proceedings, Part II*, page 744–756, Berlin, Heidelberg. Springer-Verlag.

Devlin, J., M. Chang, K. Lee, and K. Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

DiCiccio, T. J. and B. Efron. 1996. Bootstrap confidence intervals. *Statistical science*, 11(3):189–228.

Ermakova, L., T. Miller, F. Regattin, A.-G. Bosser, C. Borg, É. Mathurin, G. Le Corre, S. Araújo, R. Hannachi, J. Boccou, A. Digue, A. Damoy, and B. Jeanjean. 2022a. Overview of joker@clef 2022: Automatic wordplay and humour translation workshop. In A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 447–469, Cham. Springer International Publishing.

Ermakova, L., T. Miller, F. Regattin, A.-G. Bosser, C. Borg, É. Mathurin, G. Le Corre, S. Araújo, R. Hannachi, J. Boccou, et al. 2022b. Overview of joker@ clef 2022: Automatic wordplay and humour translation workshop. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 447–469. Springer.

Ghanem, B., J. Karoui, F. Benamara, P. Rosso, and V. Moriceau. 2020. Irony detection in a multilingual context. In *European Conference on Information Retrieval*, pages 141–149. Springer.

Grover, K. and T. Goel. 2021. Haha@iberlef2021: Humor analysis using ensembles of simple transformers. In M. Montes, P. Rosso, J. Gonzalo, M. E. Aragón, R. Agerri, M. Á. Á. Carmona, E. A. Mellado, J. Carrillo-de-Albornoz, L. Chiruzzo, L. A. de Freitas, H. Gómez-Adorno, Y. Gutiérrez, S. M. J. Zafra, S. Lima, F. M. P. del Arco, and M. Taulé, editors, *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*, volume 2943 of *CEUR Workshop Proceedings*, pages 883–890. CEUR-WS.org.

Hinton, G., N. Srivastava, and K. Swersky. 2012. Lecture 6a overview of mini–batch gradient descent. *Coursera Lecture slides https://class. coursera. org/neuralnets-2012-001/lecture,[Online.*

Hossain, N., J. Krumm, M. Gamon, and

H. Kautz. 2020. SemEval-2020 task 7: Assessing humor in edited news headlines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 746–758, Barcelona (online), December. International Committee for Computational Linguistics.

Howard, J. and S. Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July. Association for Computational Linguistics.

Low, P. A. 2011. Translating jokes and puns. *Perspectives*, 19(1):59–70.

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

Mihalcea, R. and C. Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 531–538, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Miller, T. 2019. The punster's amanuensis: The proper place of humans and machines in the translation of wordplay. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 57–65, Varna, Bulgaria, September. Incoma Ltd., Shoumen, Bulgaria.

Paolillo, J. C. 2007. How Much Multilingualism?: Language Diversity on the Internet. In *The Multilingual Internet: Language, Culture, and Communication Online*. Oxford University Press, 05.

Popa, D.-E. 2005. Jokes and translation. *Perspectives: Studies in Translatology*, 13(1):48–57.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing @ NeurIPS 2019*.

Savage, B. M., H. L. Lujan, R. R. Thipparthi, and S. E. DiCarlo. 2017. Humor, laughter, learning, and health! a brief review. *Advances in physiology education*.

Song, K., K. M. Williams, D. L. Schallert, and A. A. Pruitt. 2021. Humor in multimodal language use: Students' response to a dialogic, social-networking online assignment. *Linguistics and Education*, 63:100903.

Subies, G. G., D. B. Sánchez, and A. Vaca. 2021. BERT and SHAP for humor analysis based on human annotation. In M. Montes, P. Rosso, J. Gonzalo, M. E. Aragón, R. Agerri, M. Á. Á. Carmona, E. Á. Mellado, J. Carrillo-de-Albornoz, L. Chiruzzo, L. A. de Freitas, H. Gómez-Adorno, Y. Gutiérrez, S. M. J. Zafra, S. Lima, F. M. P. del Arco, and M. Taulé, editors, *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*, volume 2943 of *CEUR Workshop Proceedings*, pages 821–828. CEUR-WS.org.

Tomás, D., R. Ortega-Bueno, G. Zhang, P. Rosso, and R. Schifanella. 2022. Transformer-based models for multimodal irony detection. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–12.

Tsakona, V. 2017. Genres of humor. In *The Routledge handbook of language and humor*. Routledge, pages 489–503.

Vásquez, C. and E. Aslan. 2021. "cats be outside, how about meow": Multimodal humor and creativity in an internet meme. *Journal of Pragmatics*, 171:101–117.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, M., H. Yang, Y. Qin, S. Sun, and Y. Deng. 2020. Unified humor detection based on sentence-pair augmentation and transfer learning. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 53–59, Lisboa, Portugal, November. European Association for Machine Translation.

Wang, Z., S. Hale, D. I. Adelani, P. Grabowicz, T. Hartman, F. Flöck, and D. Jurgens. 2019. Demographic inference and representative population estimates from multilingual social media data. In *The World Wide Web Conference*, pages 2056–2067.

Wu, J., H. Lin, L. Yang, and B. Xu. 2021. Mumor: A multimodal dataset for humor detection in conversations. In *Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part I*, page 619–627, Berlin, Heidelberg. Springer-Verlag.

Yang, Z., L. Ai, and J. Hirschberg. 2019. Multimodal indicators of humor in videos. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 538–543.

Zabalbeascoa, P. 2005. Humor and translation - an interdiscipline. *Humor-International Journal of Humor Research*, 18(2):185–207.