LLI-UAM Team at FinancES 2023: Noise, Data Augmentation and Hallucinations

Jordi Porta-Zamorano[1,†], Yanco Torterolo[1] and Antonio Moreno-Sandoval[1]

[1]*Laboratorio de Lingüística Informática, Universidad Autónoma de Madrid, Cantoblanco, 28049, Madrid, Spain.*

Abstract

This paper describes the T5-based system developed for FinancES 2023 Shared Task by the Laboratorio de Lingüística Informática at UAM. The LLI-UAM system achieved a good ranking in all the tasks. The paper also describes some noise and data augmentation or hallucination mitigation experiments. In particular, we used corrected versions of the datasets to evaluate the impact of noise. Moreover, ChatGPT was utilised to augment the data and improve accuracy in tagging. We also describe the presence of hallucinations. Ultimately, we identify the best model for each task and draw conclusions based on our findings.

Keywords

Data augmentation, ChatGPT, noise, hallucinations, mT5, FinancES shared task

1. Introduction

Financial narrative processing has become an emerging field within NLP in recent years. In particular, FNP workshops (http://wp.lancs.ac.uk/cfie/) have spread the use of different approaches (ML and Deep learning methods especially) in tasks such as summarisation, concept and term extraction, document structure analysis, cause-effect relation detection, readability metrics and also sentiment analysis. Most of the work has focused on English, the language *par excellence* of finance, but it has recently been applied to Spanish ([1]). Within the sentiment analysis in financial Spanish, we can mention the recent works of [2] and [3]. However, both have focused on CEO letters to shareholders within annual reports.

In contrast, FinancES 2023 has as its data source those extracted from financial news headlines collected from specialised digital newspapers. The purpose of FinancES 2023 is to host a shared-task competition for evaluating competing targeted SA systems in Spanish. FinancES has been organised within the 2023 edition of the Iberian Languages Evaluation Forum (IberLEF 2023) [4].

The FinancES shared task [5] aims to investigate targeted sentiment analysis within the financial domain, encompassing diverse perspectives: (1) the economic target of the news

IberLEF 2023, September 2023, Jaén, Spain

*Corresponding author.

jordi.porta@uam.es (J. Porta-Zamorano); yanco.torterolo@inv.uam.es (Y. Torterolo); antonio.msandoval@uam.es (A. Moreno-Sandoval)

http://www.lllf.uam.es/ (J. Porta-Zamorano)

0000-0001-5620-4916 (J. Porta-Zamorano); 0000-0002-3688-3293 (Y. Torterolo); 0000-0002-9029-2216 (A. Moreno-Sandoval)

Ranking	Team	F1 Task 1	F1 Target	F1 Target Sentiment
1	abc111	0.792244 (1)	0.877137 (1)	0.707350 (4)
2	LLI-UAM	0.792172 (2)	0.852179 (3)	0.732164 (1)
3	ABCD Team	0.782175 (3)	0.854511 (2)	0.709838 (3)
4	SINAI	0.778002 (4)	0.838174 (4)	0.717829 (2)
5	AnkitSinghRaikuni	0.554211 (5)	0.575360 (6)	0.533062 (7)
6	UTB-NLP	0.529229 (6)	0.410079 (8)	0.648379 (5)
7	NLP_URJC	0.514414 (7)	0.606773 (5)	0.422055 (9)
8	BASELINE	0.498107 (8)	0.428393 (7)	0.567822 (6)
9	mario.pv	0.276926 (9)	0.106326 (9)	0.447526 (8)
10	UNAM Text Mining	0.134680 (10)	0.086643 (10)	0.182717 (10)
11	fanchuyi	0.000000 (11)	0.000000 (11)	0.000000 (11)

(a) Task 1: Financial targeted sentiment analysis.

Ranking	Team	F1 Task 2	F1 Companies Sentiment	F1 Consumers Sentiment
1	LLI-UAM	0.642349 (1)	0.592590 (1)	0.692109 (1)
2	SINAI	0.634901 (2)	0.583483 (3)	0.686320 (2)
3	ABCD Team	0.610373 (3)	0.588635 (2)	0.632111 (3)
4	abc111	0.575015 (4)	0.530284 (4)	0.619746 (4)
5	fanchuyi	0.472685 (5)	0.414230 (7)	0.531139 (5)
6	AnkitSinghRaikuni	0.457632 (6)	0.419755 (6)	0.495509 (6)
7	BASELINE	0.433783 (7)	0.384268 (8)	0.483298 (7)
8	NLP_URJC	0.425126 (8)	0.436560 (5)	0.413692 (8)
9	UNAM Text Mining	0.370396 (9)	0.345686 (9)	0.395107 (9)
10	mario.pv	0.248196 (10)	0.269267 (10)	0.227125 (10)
11	UTB-NLP	0.000000 (11)	0.000000 (11)	0.000000 (11)

(b) Task 2: Financial Sentiment Analysis at document level for companies and consumers.

Table 1
Official leaderboards for FinancES 2023 Tasks.

item; (2) the individual economic agent: companies; and (3) the individual economic agent/patient: consumers. The news item impacts the target and the economic participants, categorising positivity, negativity, or neutrality. FinancES proposes two tasks: (1) identifying the target entity in the text and determining the emotional polarity towards that target, and (2) assessing the impact of a news headline on companies and consumers regarding their stance and expressed polarity values. Our systems reached the second position for Task 1 and the first position for Task 2 in the official leaderboards, as can be seen in Table 1 for the LLI-UAM Team.

This paper outlines the system developed by the LLI-UAM team, presenting their contributions to the FinancES shared task. First, the dataset and the noise found in the examples are described (sections 2 and 3, respectively). Next, we show how data augmentation has been performed with ChatGPT. In section 5, we describe the deep learning model used. The longest part is devoted to discussing the results of the different experiments (noise, data augmentation, and

Headline Text	Target	Target Sentiment	Companies Sentiment	Consumers Sentiment
Peligroso atasco en los fondos que invierten en renovables	fondos que invierten en renovables	negative	neutral	negative
El fondo de recuperación de la UE 'va demasiado lento', según el ministro de economía francés	fondo de recuperación	positive	negative	negative
El PSOE propone un intrumento para abaratar los préstamos a las empresas similar al británico	PSOE	positive	positive	neutral
Madrid negocia con Economía para poder destinar las viviendas del 'banco malo' a desahuciados	Madrid	positive	neutral	positive

Table 2
Dataset Examples

hallucinations). We end with some reflections and proposals for future work.

2. The FinancES Dataset

The FinancES dataset and its annotation process are detailed in [5] and [6]. The dataset consists of news headlines written in Spanish, collected from digital newspapers specialised in economic and financial news from various Spanish-speaking countries. Each headline is labelled to identify the target and sentiment polarity across three dimensions: target, companies, and consumers, employing a three-class polarity value system (positive, neutral, or negative). According to [6], three organisation committee members manually annotated each headline. In cases of disagreement, the annotators engaged in discussions to resolve the matter, and if no consensus was reached, the headline was excluded. Table 2 illustrates a selection of examples from the dataset.

3. Noise in the Dataset

Annotated data holds paramount importance for training and evaluating machine learning models. Consequently, the annotations should exhibit a high level of accuracy. However, recent research has demonstrated that this is only sometimes the case, revealing a surprising number of annotation errors or inconsistencies in even widely-used datasets [7]. Since humans typically carry out dataset annotations, errors or inconsistencies are an inherent possibility. Such inaccuracies can adversely affect a model's performance, potentially leading to erroneous predictions. Although effective, rectifying these labelling errors incurs high costs and demands

substantial time investments.

Upon inspecting the dataset, we identified errors in target and polarity labels that could impact the model's performance. These errors were readily apparent, as we expected the target to be mentioned within the news headline and the polarity values to adhere to the three labels. Most of the spotted target and label errors are recurrent:

- The omission of the segment *el* within the target, e.g., **Barcó* (*Barceló*)
- Extra blanks and extra or missing quotation marks in the target or sentiment field, e.g., **telecos'* (*'telecos'*), **positive␣* (*positive*)
- Casing, e.g., **SUARA* (Suara) or **hosteLEría* (*hostelería*)
- Typographical errors in the sentiment labels, e.g., **postive* (*positive)*

Validating polarity values presents additional challenges, requiring domain knowledge in finance, and only a few labels were modified during the review process. Namely, ten were adjusted due to typographical errors, extra spaces, or characters.

Over three hundred instances from the dataset samples were corrected and handled separately.

4. Data Augmentation

Machine learning models' effectiveness and overall capabilities rely on the training data's quality, quantity, and relevance. Unfortunately, gathering enough data can be difficult and costly, resulting in a shortage of available data.

Data augmentation (DA) refers to strategies for increasing the diversity of training examples without gathering more data directly [8]. Similarly to AugGPT [9], we used ChatGPT to rephrase or paraphrase some of the training examples to enhance the training set. To this end, we designed a step-wise prompt that provides an example of rephrasing a news headline five times. In addition, the prompt includes instructions on maintaining the target's format and other labels in the generated examples since both the input and rephrased examples are in CSV.

For instance, the headline *Pemex no reemplazará a 3.000 empleados que se jubilarán este año* can be restated as:

- *La petrolera Pemex no tiene previsto sustituir a 3.000 trabajadores que se retirarán este año*
- *Pemex se niega a contratar nuevos empleados para reemplazar a los 3.000 trabajadores que se jubilarán este año*
- *La petrolera Pemex enfrenta la no renovación de 3.000 trabajadores jubilados y no tiene planes para reemplazarlos*
- *El plan de Pemex no incluye el reemplazo de los 3.000 trabajadores que se retirarán este año, lo que podría afectar su productividad*
- *Pemex anuncia que no habrá sustitución de los 3.000 empleados que se jubilarán este año, lo que genera preocupación sobre su capacidad para mantener su producción*

The last two rephrases of the previous example add a consequence or effect to the initial statement.

It is important to note that ChatGPT may occasionally fail to preserve the original intended meaning in its rephrased versions and the target. The latter is more likely to occur when the target word is capitalised at the start of a sentence (but is not a proper noun) or when it can be easily paraphrased. The headline *Restaurantes elásticos para sobrevivir* exemplifies both problems at preserving the target *Restaurantes elásticos*:

- *La adaptabilidad de los restaurantes les permite sobrevivir en situaciones cambiantes*
- *La capacidad de los restaurantes para adaptarse les ayuda a superar los desafíos*
- *Los restaurantes se vuelven flexibles para mantenerse a flote en tiempos difíciles*
- *La elasticidad de los restaurantes es clave para su supervivencia*
- *Los restaurantes demuestran su resiliencia al adaptarse a las circunstancias*

Data augmentation was performed blindly on 527 training examples, resulting in five rephrased entries from each. Additionally, we chose 250 entries that were accurately annotated from a linguistic perspective to augment similarly. As a result, we created a new dataset with 3885 examples to use for experimentation.

5. Modeling

We decided to approach all the FinancES tasks with the T5 end-to-end architecture [10]. We initially considered byT5 [11] and mT5 [12]. After conducting multiple experiments, it was determined that byT5 (a byte-based multilingual version of T5) was less effective than mT5 due to longer training times and inferior results. The mT5 model is a massively multilingual pre-trained text-to-text transformer that can be simultaneously fine-tuned on multiple downstream tasks using a task prefix or prompt.

The tasks related to target, target sentiment, company sentiment, and consumer sentiment annotations have been divided into sub-tasks. This is illustrated in Figure 1, where each annotation in the example has a different prefix indicating the specific task that needs to be performed on the headline and the expected output. The mT5 model comes in different sizes, but only small, base, and large models were chosen to experiment with since only these models fit into the single RTX 3090 24GB GPU card available for this work.

6. Experiments and Results

The data provided for training was divided into two sets: the training set and the development test set. The examples in the development set align with the ones given to participants for practice. We conducted experiments using three different versions of the original training set:

1. The original training set (T)
2. The augmented original training set (T+A)
3. The corrected training set (T')
4. The augmented corrected training set (T'+A')

Headline Text	Target	Target Sentiment	Companies Sentiment	Consumers Sentiment
Renfe afronta mañana un nuevo día de paros parciales de los maquinistas	Renfe	negative	negative	negative

(a) Original Example

Input	Output
target: Renfe afronta mañana un nuevo día de paros [...]	Renfe
target_sentiment: Renfe afronta mañana un nuevo día de paros [...]	negative
companies_sentiment: Renfe afronta mañana un nuevo día de paros [...]	negative
consumers_sentiment: Renfe afronta mañana un nuevo día de paros [...]	negative

(b) Converted Example

Figure 1: Data conversion for the mT5 model.

As the development set, we used the corrected versions for all the experiments. Throughout the training process, we used a more straightforward metric called exact match (also known as subset accuracy) instead of more complicated F1-based metrics in our multi-task framework. This metric was employed as the early-stopping criterion on the development set. The following hyperparameters, chosen tentatively, were common to all the experiments:

- learning rate: 1e-4 (constant)
- weight decay: 0.01
- batch size: 12
- optimizer: Adafactor
- epochs: 100 / patience: 10

6.1. Results on Noise and Data Augmentation

Figures 2 and 3 display scatter plots of the results for the two tasks and their sub-tasks. These figures show each scatter plot's model size, training dataset, and metrics (F1 on IOB2 tags for target and F1 for sentiments and tasks). Generally, we observe an improvement in metrics for any task and sub-task related to sentiments as the size of the model increases. For Task 1, target F1 are less clear-cut than with sentiments. For Task 2, augmented uncorrected training data obtains better results than other training sets.

6.2. Results on Hallucinations

Any language model generating content is prone to hallucinate unintended text, which can harm the system's performance [13]. Using mT5 for the tasks, we only observed hallucinations in the form of unfaithful text in target identification or fabricated targets. We categorize these mistakes as "hallucinations" within our system. They can be grouped as follows:

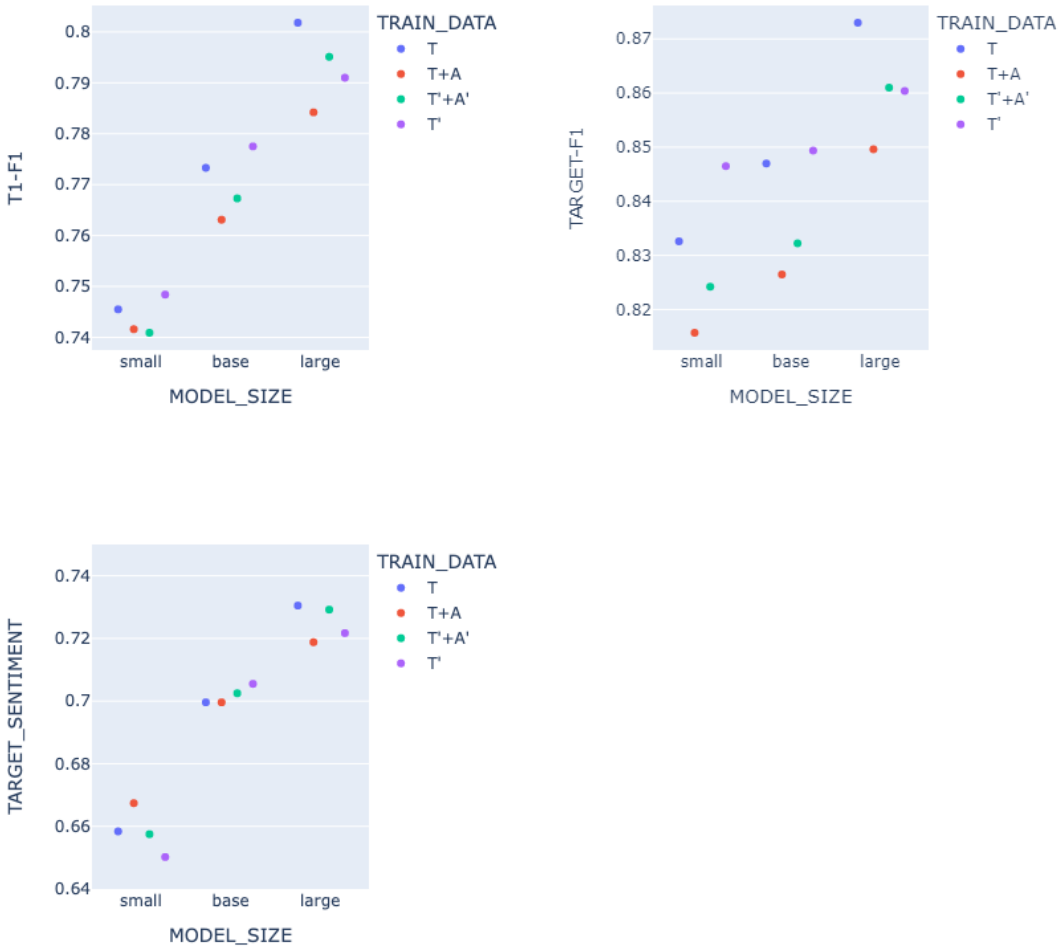

Figure 2: Task 1 results: Scatter plots of Task 1 metrics for different model sizes and training datasets.

- Typographical hallucinations. Affecting spacing: **jubilacionesforzosas y anticipadas* (*jubilaciones forzosas y anticipadas*), serial punctuation: **Santander, Sabadell BBVA y CaixaBank* (*Santander, Sabadell, BBVA y CaixaBank*); casing: **hidrógeno* (*Hidrógeno*), **Dos Heridos* (*Dos heridos*), **empresas Alicantinas* (*empresas alicantinas*), **coronavirus* (*Coronavirus*), **ministerio* (*Ministerio*), and **Unicaja Y LIBERBANK* (*Unicaja y Liberbank*); but one of the most recurrent patterns observed is the capitalized segment *le* inside a word: **TeLEfónica*, **hosteLEría*, **hosteLEros*, **TeLEcinco*, **hoteLEs*, **cadenas hoteLEras*, and **TeLEpizza*, or the segment *el*: **MerkEL*, and **ELéctricas*.
- Hallucinations inside words: **Cada hogagar* (*Cada hogar*), **Barcclays* (*Barclays*), **startup-ups* (*start-ups*), **modeloo Alzira* (*modelo Alzira*), **Tefónica* (*Telefónica*), **"inflación multipólica"* (*"inflación monopólica"*), and **marcas líders en gran consumo* (*marcas líderes en*

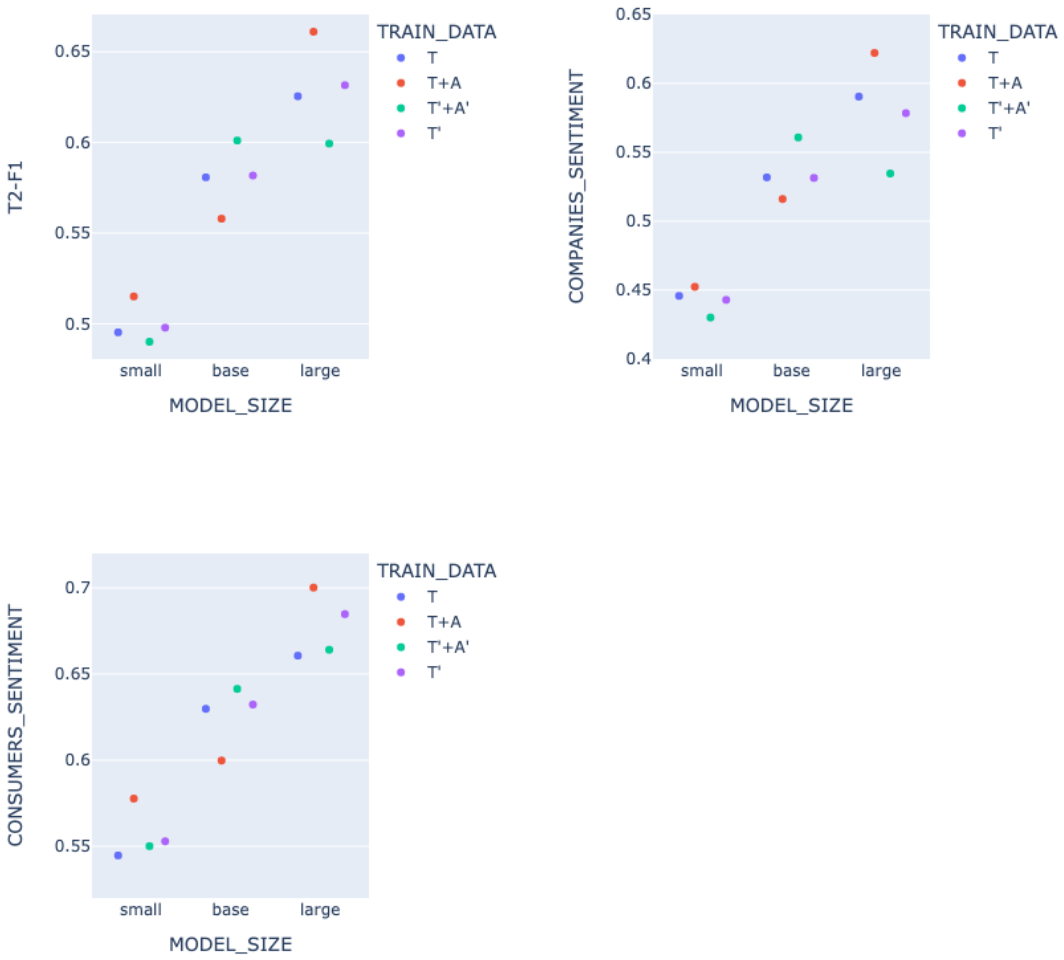

Figure 3: Task 2 results: Scatter plots of Task 2 metrics for different model sizes and training datasets.

gran consumo).

- Lexical hallucinations (some words are replaced by other somehow related): **teatral Lliure* (*teatro Lliure*), **motos minera* (*marcha minera*), **Mar del Norte* (*Mar del Sur*), **Bolsa de Buenos Argentina* (*Bolsa de Buenos Aires*), **Argentina de regulación* (*Aires de regulación*), and **web* (*Internet*).

However, Typographical hallucinations, like *TeLEcinco* or *TeLEpizza*, are not considered genuine hallucinations because they replicate the same errors as found in the sample datasets, such as *TeLEfónica*. These can be more accurately explained as an instance of noise amplification or error overfitting.

In order to deal with hallucinations in the post-processing stage, it is necessary to anchor

Model Size	Train. Set	Uncorr. F1	Corr. F1	F1 Diff.
small	T	0.8326	0.8396	+0.0070
small	T'	0.8465	0.8535	+0.0070
small	T+A	0.8157	0.8219	+0.0062
small	T'+A'	0.8242	0.8300	+0.0058
base	T	0.8470	0.8551	+0.0081
base	T'	0.8494	0.8526	+0.0032
base	T+A	0.8265	0.8287	+0.0022
base	T'+A'	0.8322	0.8398	+0.0076
large	T	0.8730	0.8732	+0.0002
large	T'	0.8604	0.8616	+0.0012
large	T+A	0.8496	0.8526	+0.0030
large	T'+A'	0.8610	0.8620	+0.0010

Table 3
Target F1

the target predictions to the headline text. This can be achieved by identifying all headline variations, completing partial words, and using a limited form of string matching.

Table 3 displays the results for post-correction of the target, showing the F1s of the original and corrected versions of the systems' output. The difference column indicates a slight improvement in the corrected versions regardless of the model's size or training set used. However, the improvement decreases on average as the model's size increases.

6.3. Best Systems

Finally, because no single training dataset fits all tasks, the best-performing systems for each of the FinancES 2023 Tasks are the following:

- Task 1:
  - Model: mT5-large
  - Training set: Uncorrected training set (T)
  - Task F1: 0.8019
  - Target F1: 0.8732
  - Target sentiment F1: 0.7305

	Precision	Recall	F1-score	Support
negative	0.771822	0.840000	0.804469	600
neutral	0.880435	0.395122	0.545455	205
positive	0.812785	0.872549	0.841608	816
macro avg	0.821681	0.702557	0.730510	1621

- Task 2:

- Model: mT5-large
- Training set: Augmented uncorrected training set (T+A)
- Task F1: 0.6611
- Companies sentiment F1: 0.6220

	Precision	Recall	F1-score	Support
negative	0.407339	0.804348	0.540804	276
neutral	0.715375	0.684915	0.699814	822
positive	0.878893	0.485660	0.625616	523
macro avg	0.667202	0.658307	0.622078	1621

- Consumers sentiment F1: 0.7002

	Precision	Recall	F1-score	Support
negative	0.588068	0.781132	0.670989	265
neutral	0.751693	0.829390	0.788632	803
positive	0.783290	0.542495	0.641026	553
macro avg	0.707684	0.717672	0.700216	1621

7. Conclusions and Future Work

While correcting hallucinations and accessing the large model are beneficial, it still needs to be determined how augmenting or correcting the training set will improve the FinancES shared tasks. According to [8], a plausible hypothesis suggests that adding more data may not necessarily improve the performance of large pre-trained transformers when working on tasks that already have sufficient representation in the pretraining data. Whether or not this hypothesis applies to the FinancES tasks is left as future work.

Acknowledgements

This publication is part of the project "Computational linguistic methods for readability and simplification of financial narratives." CLARA-FINT (PID2020-116001RB-C31), funded by the Spanish Ministry of Science and Innovation and the State Research Agency.

References

[1] A. Moreno-Sandoval (Ed.), Financial Narrative Processing in Spanish, Tirant lo Blanch, Valencia, 2021.

[2] A. Moreno-Sandoval, A. Gisbert, P. A. Haya, M. Guerrero, H. Montoro, Tone Analysis in Spanish Financial Reporting Narratives, in: Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019) NoDaLiDa, Association for Computational Linguistics, Online, 2019. URL: https://aclanthology.org/W19-6406.pdf.

[3] N. Bel, G. Bracons, S. Anderberg, Finding Evidence of Fraudster Companies in the CEO's Letter to Shareholders with Sentiment Analysis, Information 12 (2021).
[4] S. M. Jiménez-Zafra, F. Rangel, M. Montes-y Gómez, Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS.org, 2023.
[5] J. A. García-Díaz, Almela, F. García-Sánchez, G. Alcaráz Mármol, M. J. Marín-Pérez, R. Valencia-García, Overview of FinancES 2023: Financial Targeted Sentiment Analysis in Spanish, Procesamiento del Lenguaje Natural 71 (2023).
[6] P. Ronghao, J. A. García-Díaz, F. García-Sánchez, R. Valencia-García, Evaluation of transformer models for financial targeted sentiment analysis in Spanish, PeerJ Computer Science 9 (2023).
[7] C. G. Northcutt, A. Athalye, J. Mueller, Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks, 2021. arXiv:2103.14749.
[8] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, E. Hovy, A Survey of Data Augmentation Approaches for NLP, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021.
[9] H. Dai, Z. Liu, W. Liao, X. Huang, Y. Cao, Z. Wu, L. Zhao, S. Xu, W. Liu, N. Liu, S. Li, D. Zhu, H. Cai, L. Sun, Q. Li, D. Shen, T. Liu, X. Li, AugGPT: Leveraging ChatGPT for Text Data Augmentation, 2023. arXiv:2302.13007.
[10] J. Ni, G. Hernandez Abrego, N. Constant, J. Ma, K. Hall, D. Cer, Y. Yang, Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models, in: Findings of the Association for Computational Linguistics: ACL 2022, Association for Computational Linguistics, Dublin, Ireland, 2022.
[11] L. Xue, A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts, C. Raffel, ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models, Transactions of the Association for Computational Linguistics 10 (2022).
[12] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021.
[13] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of Hallucination in Natural Language Generation, ACM Computing Surveys 55 (2023) 1–38.