# Team ITST at FinancES 2023: A Psycholinguistic-based Sentiment Analysis Approach

María del Pilar Salas-Zárate [1,†] and Mario Andrés Paredes-Valverde [1,*,†]

[1] *Tecnológico Nacional de México/I.T.S. Teziutlán, Fracción l y ll SN, 73960 Teziutlán, Puebla, Mexico*

**Abstract**

This paper describes the participation of the ITST team in the FinancES 2023 shared task on financial targeted sentiment analysis in Spanish. This paper proposes a sentiment analysis approach based on psycholinguistic features which are obtained through the LIWC tool. Since the features provided by LIWC are many, the use of a feature selection technique based on Rough Set Theory and Information Gain is proposed to eliminate irrelevant features and thus improve the performance of the generated model. With respect to feature selection, a significant difference can be seen in terms of the LIWC categories selected for determining the sentiment polarity of each news headline towards both consumers and companies.

**Keywords**

Natural Language Processing, Psycholinguistics features, LIWC, Feature selection

## 1. Introduction

Sentiment analysis has become a relevant technology because it allows us to understand people's opinions regarding various topics such as product preferences, marketing campaigns, politics, among others. Today, there is a wealth of information published on the web related to the financial domain. This represents a great opportunity to understand public opinion and generate solutions based on sentiment analysis to support decision making. However, sentiment analysis in the financial context represents a major challenge because the language used in this context is inherently complex since financial terms refer to an underlying social, economic, and legal context (Milne & Chisholm, 2013).

This paper concerns our participation at FinancES 2023 Task (García-Díaz et al., 2023) which belongs to the IberLEF 2023 (Jiménez-Zafra et al., 2023). This work seeks to determine if psycholinguistic features can be used to improve the performance of sentiment analysis for financial domain. For this purpose, the financial dataset was analyzed using the Spanish version of the Linguistic Inquiry and Word Count (LIWC) program (Tausczik & Pennebaker, 2009). LIWC counts words in psychologically meaningful categories such as cognitive process, positive emotion, negative emotions, discrepancy, negation, certainty, among others. Also, we implemented a feature selection process based on Rough Set Theory and Information Gain which aims to improve the model performance by eliminating irrelevant categories thus allowing the model focuses on the most important information, which can result in a better generalization capability.

Next section describes the developed strategies for identifying the main economic target from news headlines as well as for determining the sentiment polarity of each news headline towards both companies and consumers. Finally, under final remarks, we discuss our results and a proposal for future work.

IberLEF 2023, September 2023, Jaén, Spain
*Corresponding author
[†] These authors contributed equally
EMAIL: maria.sz@teziutlan.tecnm.mx (M.P. Salas-Zárate); mario.pv@teziutlan.tecnm.mx (M.A. Paredes-Valverde)
ORCID0000-0003-1818-3434 (M.P. Salas-Zárate); 0000-0001-9508-9818 (M.A. Paredes-Valverde)

IberLEF 2023, September 2023, Jaén, Spain.

## 2. Developed strategies

### 2.1. Task 1: Financial targeted sentiment analysis

#### 2.1.1. Determining the sentiment polarity

Figure 1 shows the flow diagram of the sentiment analysis process proposed in this paper, which consists of three main phases, namely, psycholinguistic feature extraction, feature selection, and modeling. Next, a brief description of each of these phases is provided.

- Dataset. The dataset used in this work was the one developed by (Pan et al., 2023), which contains tweets in the financial context and news headlines in Spanish.
- Psycholinguistic feature extraction. This phase aims to better understand the emotions, attitudes, themes, and subjectivity present in texts, in this case, in the opinions related to the financial context. For this purpose, the LIWC tool was used, which uses a linguistic dictionary to analyze the emotional, cognitive, and linguistic content of a text. In other words, LIWC makes it possible to identify and measure the emotions and attitudes expressed by users through text. Among other things, it can detect emotional tone, happiness, sadness, anger, or fear, and thus provide information about the overall sentiment expressed in the text.
- Feature selection. LIWC provides a total of 72 categories grouped in 5 dimensions which are linguistic process, psychological process, personal concerns, spoken categories, and punctuation marks. As can be seen, LIWC provides a large set of categories for each of the texts analyzed. In this sense, a feature selection process was developed to identify and select the most relevant and useful categories to build the model. This phase is of great importance because not all categories provided by LIWC contribute in the same way to the performance of the model. In summary, this phase aims to improve the model performance since, by eliminating irrelevant categories the model focuses on the most important information, which can result in a better generalization capability. Finally, it is important to mention that a hybrid approach to feature selection based on Rough Set Theory (RST) (Pawlak, 1982)and Information Gain (IG) (Arafat et al., 2014)was used in this phase. This feature selection approach was implemented in (Salas-Zárate & Paredes-Valverde, 2016)obtaining encouraging results.
- Modeling. In this stage, different models were built and trained using SVM, BayesNet and J48 algorithms. The generated models were evaluated with a different data set than the one used for the training phase. SVM obtained the best results and was therefore selected for the shared task.

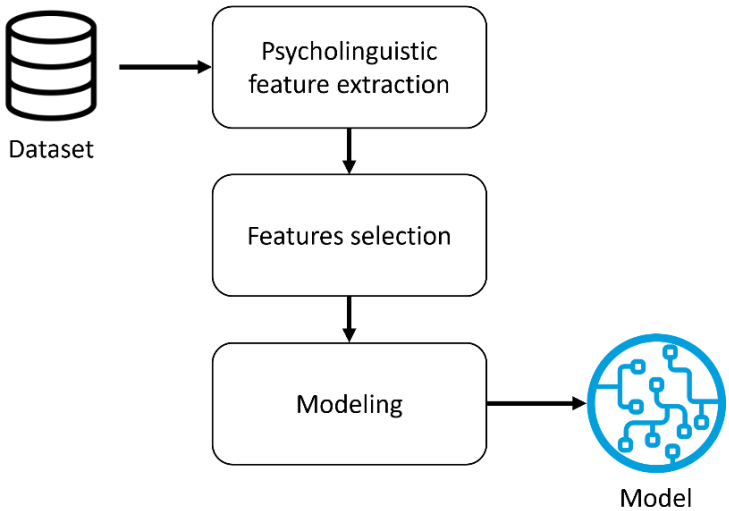


**Figure 1**: Flow diagram of the psycholinguistic-based sentiment analysis approach.

The process described above was implemented for both this task (Determining the sentiment polarity) and task 2 (Financial sentiment analysis at document level for companies and consumers). However, for each of the tasks, different categories were selected. With respect to task 1, the result of

the feature selection process was 45 selected categories which are shown in Table 1. As can be seen, the dimensions of linguistic process and psychological process contain a greater number of selected categories. This may be because such categories are more related to the understanding of emotions such as anxiety, anger, sadness, positive, negative, among others, which are present in the expression of opinions.

**Table 1**
LIWC categories selected for sentiment analysis.

| Set | Categories |
|---|---|
| Linguistic process (LP) | WC, WPS, BigWords, Dic, Funct, TotPron, PronPer, ElElla, Ellos, PronImp, Articulo, Verbos, Adverb, Prepos, Conjunc, Negacio, Cuantif, Numeros, informal |
| Psychological process (PP) | Social, Afect, EmoPos, EmoNeg, Ansiedad, Enfado, Triste, MecCog, Insight, Causa, Inhib, Incl, Excl, Ver, Ingerir, Relativ, Movim, Espacio |
| Personal concerns (PC) | Trabajo, Logro, Dinero, Muerte |
| Spoken categories (SC) | - |
| Punctuation marks (PM) | AllPunc, Period, Comma, OtherP |

Figure 2 shows the 10 most discriminative psycholinguistic features for the sentiment analysis of task 1. As can be seen, among this group we find a higher number of features belonging to the linguistic process group with 6 while the rest corresponds to psychological process (2) and punctuation marks (2).

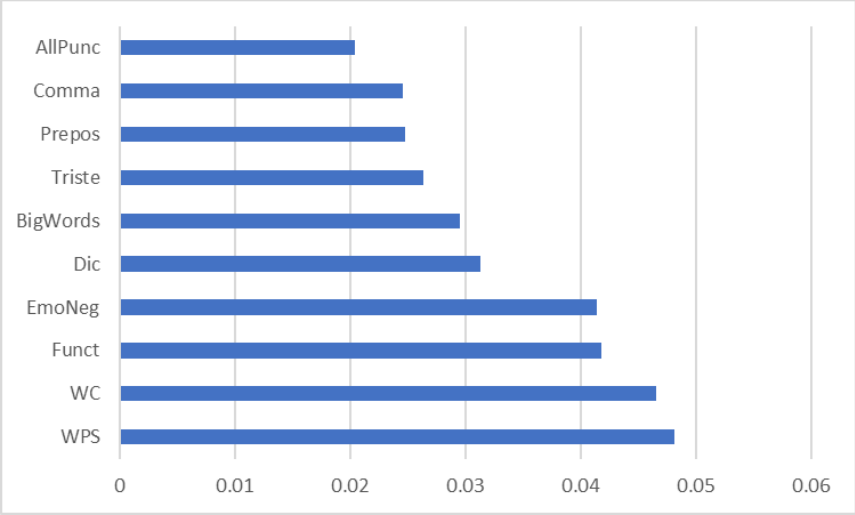


**Figure 2**. Top 10 features for Task 1 sentiment analysis.

## 2.2. Task 2: Financial Sentiment Analysis at document level for companies and consumers

### 2.2.1. Determining the sentiment polarity of news headlines towards companies

As mentioned above, Task 2 employed the process shown in Figure 1 Specifically, for the case of sentiment analysis for companies, the feature selection process resulted in a total of 23 selected LIWC categories (see Table 2). Again, the linguistic process (10) and psychological process (10) dimensions contain the largest number of categories. With respect to the psychological process, the categories of negative emotions (EmoNeg), anger (Enfado) and anxiety (Ansiedad) can be highlighted as the most discriminating characteristics in news headlines towards companies.

**Table 2**
LIWC categories selected for sentiment polarity detection for companies.

| Set | Categories |
|---|---|
| Linguistic process (LP) | Articulo, Funct, WPS, WC, Dic, BigWords, Numeros, Negacio, Prepos, VosUtds |
| Psychological process (PP) | EmoNeg, Enfado, Ansiedad, Social, Afect, Tiempo, Humanos, Inhib |
| Personal concerns (PC) | Muerte, Trabajo |
| Spoken categories (SC) | |
| Punctuation marks (PM) | Comma, AllPunc, Period |

Figure 3 shows the 10 most discriminating psycholinguistic features for the sentiment analysis of news headlines towards companies. In this case, there was a higher number of psycholinguistic features from the linguistic process category with 5 followed by the linguistic process category with 4. It should be highlighted the fact that features such as negative emotion (EmoNeg) and anger (Enfado) are more discriminative, which is related to the type of emotions expressed by users.

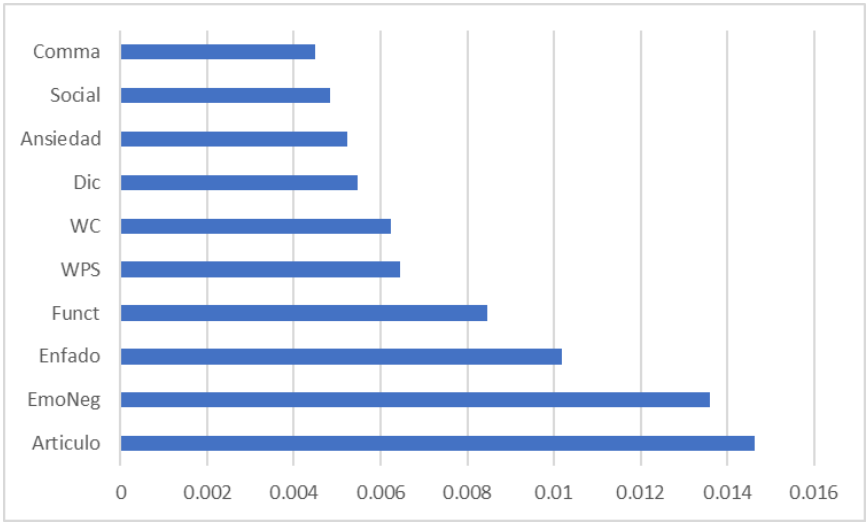


**Figure 3**. Top 10 features for Task 2 sentiment analysis of news headlines towards companies.

## 2.2.2. Determining the sentiment polarity of new headlines towards consumers

Table 3 shows the set of features selected for sentiment analysis of news headlines toward consumers. Out of the 27 categories selected there is a higher number for the linguistic (15) and psychological (10) process dimensions. Unlike the categories selected for the previous section, in this subtask the linguistic process dimension presents a larger number of categories, where the categories BigWords, ElElla, and TotPron represent some of the most relevant for the generation of the model.

**Table 3**
LIWC categories selected for sentiment polarity detection for consumers.

| Set | Categories |
|---|---|
| Linguistic process (LP) | Articulo, WPS, WC, Dic, Funct, Numeros, informal, Futuro, Conjunc, BigWords, PronImp, ElElla, PronPer, TotPron, Cuantif |
| Psychological process (PP) | Triste, Tiempo, Ansiedad, Inhib, Relativ, EmoNeg, MecCog |
| Personal concerns (PC) | Hogar, Dinero, Trabajo |
| Spoken categories (SC) | - |
| Punctuation marks (PM) | Comma, AllPunc |

Finally, Figure 4 shows the Top 10 features for Task 2 sentiment analysis of news headlines towards consumers. As can be seen, there is a great difference between the characteristics selected for the analysis of news headline sentiments towards companies and consumers. One of the main differences lies in the fact that psycholinguistic features related to negative emotions such as negative emotion (EmoNeg) and anger (Anger) that were selected for the case of companies do not appear in the case of consumers. For the latter case there are a greater number of features belonging to the category of linguistic process with 6.

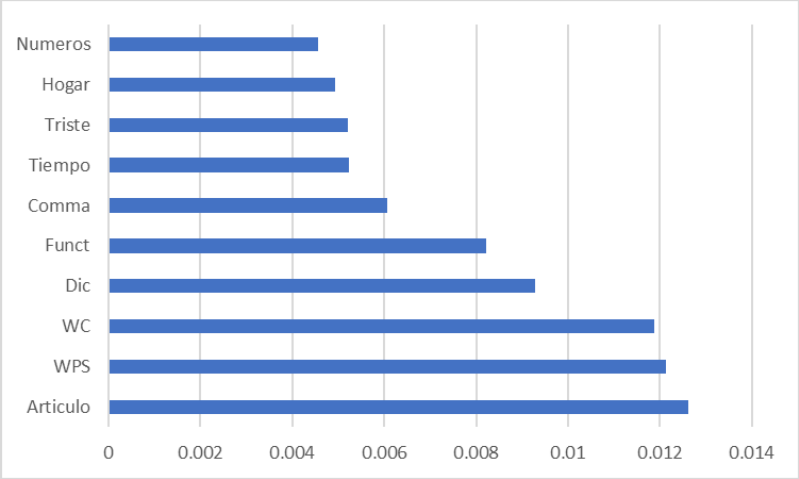


**Figure 4.** Top 10 features for Task 2 sentiment analysis of news headlines towards consumers.

## 2.3. Results

Figure 5 shows the confusion matrix corresponding to the sentiment analysis of Task 1 where it can be seen that the proposed approach is efficient in identifying positive opinions with an F-measure value of 0.74; however, in the case of neutral opinions it fails outright by not being able to identify any as such.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| negative | 0.441176 | 0.294118 | 0.352941 | 51 |
| neutral | 0.000000 | 0.000000 | 0.000000 | 9 |
| positive | 0.674074 | 0.834862 | 0.745902 | 109 |
| accuracy | | | 0.627219 | 169 |
| macro avg | 0.371750 | 0.376327 | 0.366281 | 169 |
| weighted avg | 0.567894 | 0.627219 | 0.587593 | 169 |

**Figure 5.** Task 1 sentiment analysis confusion matrix.

Table 4 shows the results obtained in the evaluation data set. As can be seen, the F1-score values obtained for each of the tasks are relatively low; however, we believe that the approach based on psycholinguistic features should not be discarded; on the contrary, it could contribute to the improvement of the accuracy of other approaches focused on financial domain since tools such as LIWC have been successfully implemented in contexts such as ecommerce(Olagunju et al., 2020), education (Geng et al., 2020), among others.

**Table 4**
Results of our approach in the evaluation dataset.

| Task | F1 score |
|---|---|
| Task 1 – Determining the sentiment polarity | 0.447526 |
| Task 2 - Determining the sentiment polarity for companies | 0.269267 |
| Task 2 - Determining the sentiment polarity for consumers | 0.227125 |

## 3. Final remarks

This paper presented a sentiment polarity detection approach for IBERLEF 2023 Task - FinancES shared task on financial targeted sentiment analysis in Spanish. For polarity detection, an approach was proposed that extracts psycholinguistic features from the opinions through the LIWC tool. In addition, a feature selection process was performed to identify and select the most relevant variables for the construction of the classifier model. As could be seen, the feature selection process produced different results for each of the targets (companies and consumers), which gives some clues as to the difference in the use of language for both cases. For future work, we intend to investigate transformers-based methods for sentiment polarity detection.

## 4. Acknowledgements

We are grateful to the Tecnológico Nacional de Mexico (TecNM, by its Spanish acronym) for supporting this work. This research was also sponsored by Mexico's National Council of Humanities, Sciences and Technologies (CONAHCYT).

## 5. References

Arafat, H., Elawady, R. M., Barakat, S., & Elrashidy, N. M. (2014). Different feature selection for sentiment classification. *International Journal of Information Science and Intelligent System*, *1*(3), 137–150.

García-Díaz, J. A., Almela, Á., García-Sánchez, F., Alcaráz Mármol, G., Marín-Pérez, M. J., & Valencia-García, R. (2023). Overview of FinancES 2023: Financial Targeted Sentiment Analysis in Spanish. *Procesamiento Del Lenguaje Natural*, *71*(0).

Geng, S., Niu, B., Feng, Y., & Huang, M. (2020). Understanding the focal points and sentiment of learners in MOOC reviews: A machine learning and SC-LIWC-based approach. *British Journal of Educational Technology*, *51*(5), 1785–1803. https://doi.org/10.1111/BJET.12999

Jiménez-Zafra, S. M., Rangel, F., & Montes-y-Gómez, M. (2023). Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages. *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*.

Milne, A., & Chisholm, M. (2013). The Prospects for Common Financial Language in Wholesale Financial Services. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.2325362

Olagunju, T., Oyebode, O., & Orji, R. (2020). Exploring Key Issues Affecting African Mobile eCommerce Applications Using Sentiment and Thematic Analysis. *IEEE Access*, *8*, 114475–114486. https://doi.org/10.1109/ACCESS.2020.3000093

Pan, R., García-Díaz, J. A., Garcia-Sanchez, F., & Valencia-García, R. (2023). Evaluation of transformer models for financial targeted sentiment analysis in Spanish. *PeerJ Computer Science*, *9*, e1377. https://doi.org/10.7717/PEERJ-CS.1377

Pawlak, Z. (1982). Rough sets. *International Journal of Computer & Information Sciences*, *11*(5), 341–356. https://doi.org/10.1007/BF01001956/METRICS

Salas-Zárate, M. del P., & Paredes-Valverde, M. A. (2016). Sentiment Classification of Spanish Reviews: An Approach based on Feature Selection and Machine Learning Methods. *J. Univers. Comput. Sci.*, *22*(5), 691–708.

Tausczik, Y. R., & Pennebaker, J. W. (2009). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Http://Dx.Doi.Org/10.1177/0261927X09351676*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676