Text-Based Matching of News

César Cabezas
Facultad de Ingeniería
Universidad del Pacífico
Lima, Perú
cm.cabezasg@alum.up.edu.pe

Franz Figueroa
Facultad de Ingeniería
Universidad del Pacífico
Lima, Perú
f.figueroag@alum.up.edu.pe

Gonzalo Alvis
Facultad de Ingeniería
Universidad del Pacífico
Lima, Perú
gr.alvisb@alum.up.edu.pe

I. Introduction

With the constant growth of digitized information, the coverage of news events by various media has gained significant relevance. In 2022, the revenues generated by the digital press amounted to approximately US$ 36.5 billion worldwide [1]. This has also implied an increase in the number of sources of dissemination and an easier access to their content. However, this has raised concerns about the quality and impartiality of what is presented. The influence of political agendas and ideological preferences can bias media coverage and affect the public's perception of different events. Consequently, the need arises to identify discrepancies in event coverage across different media outlets, such as newspapers or magazines. To tackle this challenge, a systematic analysis is imperative to discern whether a news article aligns with a specific agenda, thereby recognizing bias in the text or the author. However, the initial step towards this goal necessitates matching two or more news items covering the same event.

For this task, we need to retrieve information from online media outlets and employ a method to calculate the similarity between different attributes of news articles. News processing can be approached as a Semantic Text Matching problem, which estimates the similarity between texts. However, the difficulty arises when dealing with very long texts, as they can fall into complex linguistic structures [2]. Our primary objective is to build a classifier model to determine if two news cover the same event. While there exists previous literature on this topic, most of it focuses on news articles in English or other languages. Therefore, for the scope of this project, we will prioritize Spanish news articles.

Regarding data retrieval, Web Scraping has become a highly relevant technique in recent times. It is a process for automatic data extraction from websites using specialized software [3]. Web scraping allows us to efficiently collect and structure data from various online sources, enabling us to gather a comprehensive dataset of Spanish news articles for our analysis. According to this, we have extracted from different Peruvian news outlets characteristics such as title, description and body. We will analyze lexical matching methods to obtain similarity scores to show the related articles from different media outlets. Ultimately, this study aims to develop an effective method for identifying the best 5 news that match to another news.

II. Literature Review

Various approaches have been developed to address the topic of text matching and the algorithms used to achieve the stated objectives. However, it is crucial to consider the scale of the texts being compared. Although text matching algorithms have proven to be efficient in linking simple queries such as sentences or questions, it remains challenging to relate longer bodies of text [4]. With this in mind, works on keyword extraction, network elaboration, recommendation systems, and similarity measures in classification have been considered.

A. Keyword Extraction

Keyword extraction is a technique that involves identifying the most relevant words or expressions that describe the main content of a document, allowing it to be linked to other similar documents. A clear example is the study conducted by Koloski et al. [5], who proposed exploring keyword extraction in different languages using methods such as Term Frequency - Inverse Document Frequency (TF-IDF). The aim of their work was to simplify tag assignment in the media through automatic recognition of keywords in texts.

To this end, the authors conducted a comparison on various datasets in different languages to analyze the performance of keyword extraction in less common languages in academic literature. Different model combinations were evaluated: supervised models such as the Transformer-based Neural Tagger for Keyword Identification (TNT-KID) and Bidirectional Encoder Representations from Transformers (BERT), as well as the unsupervised TF-IDF model. The results of the study were varied, with better performance achieved when the three mentioned models were used complementarily. This process highlighted the existence of various methods for keyword extraction, which allow for the identification of the semantic identity of different texts.

B. Recommendation Systems

In the context of a recommendation system, the application of BERT is relevant for addressing the 'Cold Start' problem characteristic of such applications, because it transfers the linguistic knowledge learned during pre-training to the news recommendation task. BERT is a pre-trained language representation model developed by Google [6]. Unlike language models that analyze text in a single direction, BERT examines text bidirectionally, allowing it to handle various types of text with a better understanding of semantics and inference, as well as improving text segmentation and classification [7].

During pre-training, BERT learns contextual representations of words that capture useful semantic and syntactic information from natural language. Therefore, instead of starting training from scratch, the model can leverage these contextual word representations that capture useful semantic and syntactic information. However, its main limitation compared to the present work is that they only use the news titles for training their model, which could restrict the depth of the generated recommendations.

In this regard, Zhang et al. [8] also developed a news recommendation system tailored to users. They propose a BERT-based user-news matching model, called UNBERT. This model aims to predict the probability that a user will click on a new candidate news item, leveraging not only the news item's title but also incorporating existing knowledge about the user's preferences and behaviors. Tested using the Microsoft News Dataset (MIND), UNBERT demonstrated superior performance in comparison to other existing methodologies for news recommendation tasks.

C. Similarity Coefficients

In this last approach, Mozer et al. [9] compare 100 text matching methods for extensive documents. The work concludes by highlighting the use of a framework with spatial representations and distance measures for text analysis and processing. By representation, Mozer et al. refer to the structure shape and quantification of the document corpus, while distance metrics are used to measure the similarity between two documents based on the chosen representation. Thus, they recommend the following procedure: choose a text representation, define a similarity metric based on its covariance, implement a technique for document matching, evaluate the match quality, and repeat the first three steps until better performance is achieved.

In order to analyze similarity in text documents, Bafna et al. [10] propose using TF-IDF and fuzzy K-means in datasets such as NEWS 20, Reuters and research papers. Before clustering, they use TF-IDF technique to eliminate the most common terms and extracts only the most relevant terms from each corpus. With this, they calculate a cosine distance matrix and apply K-means with iterations and use silhouette coefficient to determine the optimal quantity of clusters.

In the work done by Umut et al. [11], they highlight the importance of matching news from different portals as a necessary step to model a comprehensive online news flow. For their project, they used 2049 Turkish news items from 20 domains, where they had 693 clusters each related to an event, with each cluster having 2.75 documents on average. They mention existing Deep Learning approaches to the topic, but with their limited database, they decided to use simple lexical similarity methods.

They used unsupervised models - using similarity indices such as Jaccard's similarity coefficient and others - to establish a weighted similarity coefficient between each article. The result of this procedure was a matrix of relationships between all of the articles, with a value between 0 and 1. They also used supervised models such as Random Forest, SVM and Neural Network. For this approach, they developed another database, labeling each article if they were covering the same event or not, having 1858 positive pairs and 15000 negative pairs. For the supervised models, they added the features obtained as a result of the unsupervised ones, such as similarity coefficients, which improved their performance. They concluded that lexical matching-based scores allow differentiating well if two news items match, even without a label, and that adding those coefficients to a supervised model increases its performance.

III. Methodology

A. Data Collection

For this project, we decided to extract information from different Peruvian digital news portals. To do this, we manually accessed each one to obtain information about the page structure. We then developed a Python script using requests and BeautifulSoup for data extraction.

For this presentation, we have collected articles from 2017-2018 from the newspapers ”El Comercio”, ”Correo” and ”Peru21”. Using a crawling procedure, we first extract the link of each one of the articles, and then the content from these. We extracted, in order of appearance, the kicker, headline, date, subheadline, body, and tags. We obtained a total of 216,577 articles. Regarding the thematic content or filters, we have considered all the articles. For this, we followed a two-step procedure consisting of fetching links and fetching data. With this, we got all the links to the news in the considered timeframe, and then we iterated between them to get all the data.

During the data collection process, we encountered several issues. We will be going through them and how we solved each.

1) Fetch Links: This process involved accessing the daily news pages and navigating through all the days within the established timeframe. On each page, we fetched all the html elements that contained news links, cleaned the links, and saved them in a text file for further use.

- **Dynamic Content**: Some news portals, like ”Perú21”, used JavaScript to load content dynamically, making it difficult to extract links using simple HTML parsing.
  Solution: We used Selenium, a browser automation tool, to render JavaScript and retrieve fully loaded pages. This allowed us to extract the links correctly.
- **Poor Response Time**: For dynamic content on portals like ”Perú21”, the time for all links to load was inconsistent and often long.
  Solution: We implemented a sleep time of 10 seconds in Python and divided the data collection tasks among all group members to gather the most links in the shortest time possible.
- **Execution Time**: The process for simpler pages like ”El Comercio” and ”Correo” was also time-consuming.

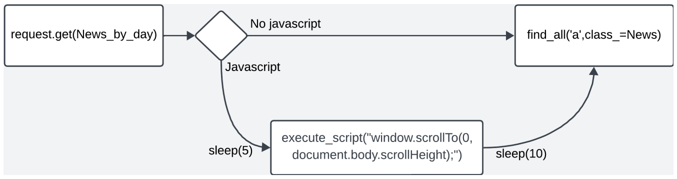
request.get(News_by_day)
No javascript
Javascript
find_all('a',class_=News)
sleep(5)
execute_script("window.scrollTo(0, document.body.scrollHeight);")
sleep(10)

Fig. 1. Fetch links process

Solution: We used `ThreadPoolExecutor` to execute multiple requests in parallel, significantly reducing the total execution time.

2) Fetch Data: For this process, we iterated through each of the previously extracted links. Since most of the news outlets shared a similar structure, we were able to use the same script for the majority of them. However, exceptions were made when the structure of certain elements differed significantly, requiring modifications to the main script to adapt to different pages.

- **Inconsistent HTML Structure**: Different articles had varying HTML structures, making it difficult to uniformly extract all required fields.
  Solution: We created a flexible parsing function that checked for multiple possible HTML structures and extracted data accordingly.
- **Missing Information**: In some cases, not all articles contained all the desired information, such as tags or subheadlines.
  Solution: We handled missing data by setting default values and ensuring our data processing pipeline could accommodate incomplete records without errors.
- **Embedded Scripts in Text**: Some articles contained embedded scripts mixed with the news content.
  Solution: We created a BeautifulSoup function to remove scripts from the HTML content.

B. Data Processing

For the development of the experiment, it is necessary to preprocess the text to remove noise, empty spaces, or other inconveniences. We can break down this stage into the following steps:

- Remove HTML characters
- Standardize words by converting them to lowercase
- Remove special characters such as accents or punctuation marks.
- Tokenize sentences through the split method of strings in python. This is the process of decomposing each sentence into smaller units, such as words and symbols. The function we used is less complex than tools like Tokenize from the NLTK library, but it allows us to process the text more efficiently, since special characters and punctuation marks, which correspond to NLTK's value proposition, have been removed before this step.
- Remove stopwords from a handmade set in spanish

We opted to retain whole words and proper nouns because of the complexities associated with working with the Spanish language, and the context of the news. Proper nouns cease to be noise when they can be talking points or words of interest in the elaboration of news groups. This is why we chose not to eliminate them. In addition, the use of Internet lemmatizers for the treatment of texts is imprecise in the Spanish language, being that the disadvantages of the use of this type of tools outweigh the advantages of the same. A first test with lemmatizers proved to be quite imprecise when reducing expressions to their minimum verbal unit.

C. Spatial Representation and Data Exploration

According to the literature reviewed, we opted for spatial representation for an initial data exploration and a review of the most relevant attributes of our news sets. For this, we will work only with the bodies of the news articles, leaving other parts such as labels, tags, or headers to reinterpret these vector representations later.

1) TF-IDF Approach: The first approach taken was through TF-IDF. TF-IDF is used to convert documents into a structured format to reflect the importance of a word to a collection of documents or corpus [10]. The components of the equation are shown in (1) and (2), where N is the total number of documents and df is the number of documents with the word, with the complete equation in (3).

$$TF(t,d) = \frac{\text{number of times } t \text{ appears in } d}{\text{total number of terms in } d} \tag{1}$$

$$IDF(N,t) = \log(\frac{N}{1+df}) \tag{2}$$

$$TF-IDF(t,d) = TF(t,d) * IDF(N,t) \tag{3}$$

Using this technique, we vectorized all the articles in the database, with the only inconvenience being the presence of empty bodies. These were associated with news that mainly contained videos or photos, representing a tiny part of the total set (less than 1% of the total articles). It was decided to eliminate these entries as they did not pose a risk to the overall integrity of the database.

The parameters considered for the TF-IDF model were as follows:

- **max_df = 0.9**: This parameter ignores terms that appear in more than 90% of the documents. It helps to remove very common words that do not provide discriminative information.
- **min_df = 295**: This parameter ignores terms that appear in less than 295 documents. It helps to remove very rare words or possible typographical errors. The number was chosen to represent the
- **use_idf = True**: This enables term re-weighting using inverse document frequency (IDF). It gives more importance to words that are less frequent in the corpus.
- **max_features = media_palabras**: This limits the vocabulary to the average number of words per corpus. It helps to control the dimensionality of the resulting vector.

- **ngram_range = (1, 3)**: This considers not only individual words but also sequences of up to 3 words. It allows capturing common phrases and expressions.

These parameters are configured to optimize the TF-IDF vectorization, balancing the capture of relevant information with the reduction of noise and management of the dimensionality of the resulting vector space. Once vectorization was done, we obtained several vectors with 153 elements. This presents a high dimensionality for spatial representation in graphs, so we opted to use dimensionality reduction techniques.

2) Word2Vec Approach: Another approach observed in the literature is Word2Vec, which we also decided to follow for the representation of news in space. Word2Vec is a method for producing vector representations of words from a high-dimensional space to a low-dimensional real vector [14]. It uses a neural network model to learn the associations of words from a large body of text. This model can be implemented using one of two architectures: Continuous Bag of Words and Skip-Gram. In this paper, we use Skip-Gram because it uses a word to predict its surrounding context. This model is trained with the corpus of the news extracted earlier. Following the same preprocessing used for TF-IDF, we used the following hyperparameters:

- Model: skipgram with negative sampling
- Context window length: 5
- Minimum word count: 5
- Training epochs: 5

3) T-SNE and K-Means: For the graphical representation of the vectors obtained through TF-IDF and Word2Vec, we used T-SNE, selected for its ability to maintain the global structure of the data, allowing us to visualize behavior patterns in the collected news. To assist in this task, we also clustered the vectors using K-Means, so that we could differentiate articles that were more similar to each other from the rest. T-SNE is a technique used for visualizing high-dimensional data in 2 or 3 dimensions [15]. It calculates similarity in high dimensions and projects it into lower dimensions to adequately reflect the original similarities. On the other hand, K-Means is a widely used clustering technique where a parameter **k** indicates the number of groups expected to be found [16].

For this process, we used the following parameters in the respective tools:

- For T-SNE, using the sklearn library, we reduced the attributes to 2 for graphical representation.
- For K-Means, using the same library, we chose a total of 10 groups to observe general classes or topics in the set.

A difficulty at this stage was the high number of groups observed through the elbow test. As shown in figure 2, the mean error level decreased constantly, seeking the ideal number of groups for 210k articles. Assuming each group is a discussion topic, this is normal, since in two years and 210 thousand articles there should be more than 150 topics present. For this reason, we decided to choose a lower number of clusters to conduct the initial exploration of the extracted news, focusing on a more general classification.

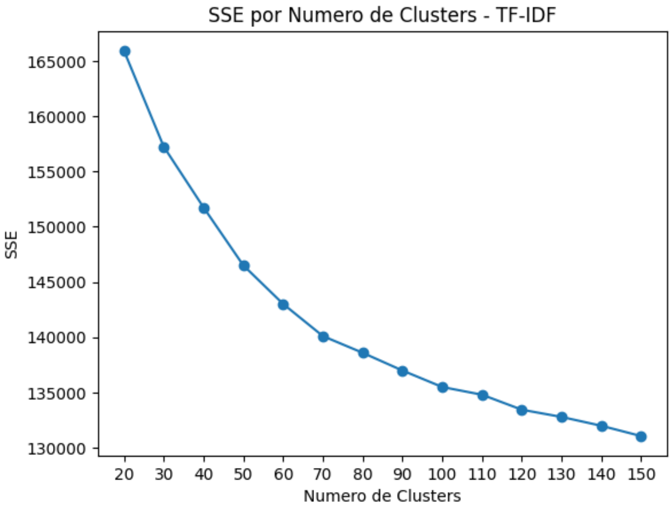

Fig. 2. Elbow test for selecting groups in the TF-IDF model

4) MLP Classification Model: Finally, we developed a Multilayer Perceptron (MLP) classification model using the sklearn library. This model consists of an artificial neural network composed of multiple layers of nodes (neurons), where each layer is fully connected to the next and is used to assign labels to data inputs, solving classification problems [17]. Each neuron processes the assigned data using an activation function. Its use in the context of news matching has been demonstrated as relevant in the work conducted by Umut et al. [11]. For our work, this model will classify articles into pairs of the same topic or not.

The labeling given on the pairs was done by hand following a date and the topics criterion.

To optimize the classification model, we compared the performance of the MLP using different vectorization techniques:

- Word2Vec
- Combination of Word2Vec with K-Means

This comparison will be made using the precision metric, which measures the proportion of correctly classified pairs over the total pairs classified as positive. It was decided to use this metric due to the primary objective of minimizing false positives to ensure that the pair predictions are reliable. It is critical to minimize false positives in the model since classifying a pair as a match when it is not could lead to erroneous analysis in bias detection between the two.

IV. Results

A. TF-IDF Analysis Results

The result of the TF-IDF analysis is shown in figure 3. In this figure, we can identify a distinguishable pattern in the data representation. Despite this, we can recognize patterns in the way groups are distributed, with areas where one group predominates over the others. In this sense, we have been able to catalog the 10 groups extracted by K-Means, labeling them in the graph and describing their main components in Table I. In this table, we identify that group 10 encompasses the noise generated by news that does not belong to any particular group.

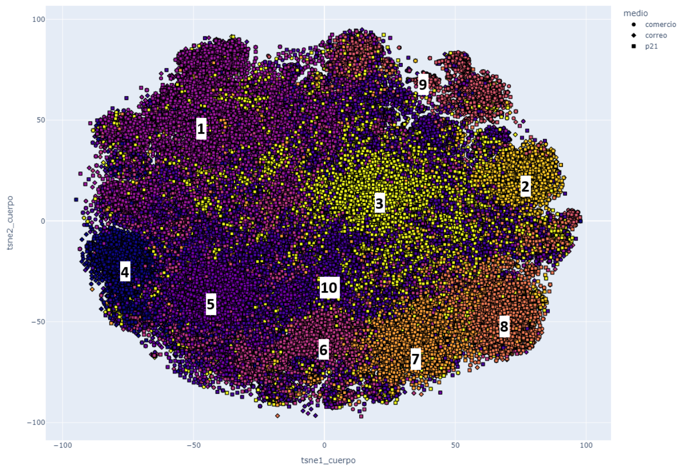

Fig. 3. General representation of data using TF-IDF and T-SNE

The labels were selected manually based on the most common words per cluster and the news present in each.

TABLE I
MOST COMMON WORDS PER CLUSTER AND LABELS

Cluster	Most common words	Label
1	investigacion, politica, fujimori	Politica
2	ver, sociales, redes, foto, cuenta	Redes Sociales
3	manera, forma, cada, hacer, tiempo	Opinion
4	distrito, region, salud, presidente	General
5	investigacion, seguridad, sospechoso	Incidentes
6	mayor, comercio, domingo, lima, san	Nacional
7	mundo, equipo, partido, peru, seleccion	Fútbol
8	mundial, equipo, partido, rusia, gran	Mundial Rusia
9	mundo, presidente, unidos, trump	Internacional
10	ciudad, parte, mujer, personas	NT

B. Word2Vec Analysis Results

The same analysis was conducted for the articles represented using Word2Vec. Figure 4 presents the result, showing less distinguishable groups than the TF-IDF analysis. However, the clusters generated using K-Means also revealed patterns that were associated with distinct topics as shown in Table II.

C. Classification Model Results

The MLP classification model was evaluated using precision as the metric. Table IV summarizes the precision results for each approach. We tested the model with Word2Vec vectors, and a combination of Word2Vec vectors with K-Means clustering. The gridsearch technique was used to find the optimal hyperparameters. For this, we took into account the following hyperparameters: hidden layer sizes, activation function, solver, learning rate and learning rate init.

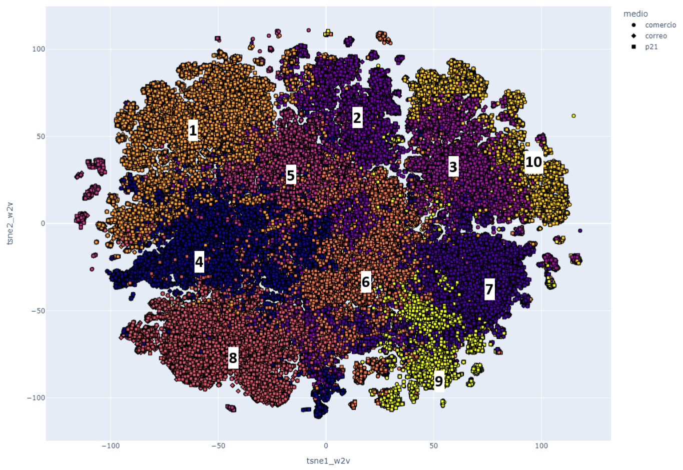

Fig. 4. General representation of data using Word2Vec and T-SNE

TABLE II
MOST COMMON WORDS PER CLUSTER AND LABELS (WORD2VEC)

Cluster	Most common words	Label
1	congreso, fujimori, corrupción, odebrecht	Corrupcion
2	presidente, trump, gobierno, venezuela	Pol. Internacional
3	selección, mundial, fútbol, partido	Deportes
4	lima, nacional, regional, salud, emergencia	Regional
5	empresas, inversión, mercado, crecimiento	Economía
6	trabajo, vida, Mundo, facebook, amor	General
7	instagram, video, redes, publicacion	Redes Sociales
8	policía, agentes, delito, menor, víctima	Crimen
9	película, serie, netflix, temporada, estreno	Entretenimiento
10	partido, copa, liga, torneo	Dep. Competitivo

TABLE III
HYPERPARAMETERS INCLUDED IN GRIDSEARCH

Hyperparameter	**Value**
`hidden_layer_sizes`	(50,), (100,), (50, 50), (100, 50), (100, 100)
`activation`	relu, tanh, logistic
`solver`	adam, sgd
`learning_rate`	constant, adaptive
`learning_rate_init`	0.001, 0.01, 0.1

V. DISCUSSION

The results of this study validate the recommendations found in the reviewed literature. The spatial representation of text is indeed an essential first step when handling extensive text bodies, such as news articles. Although both techniques chosen for spatial representation are effective, Word2Vec stands out over TF-IDF in text representation within a spatial framework. The clusters obtained through the Word2Vec model are not only more coherent in content, but this coherence also extends to their spatial distribution.

The groups in Figure 3 are filled with noise and, at times, are very dispersed or mixed, unlike the groups in Figure 4. Due to this precision in representation, we chose to base our classification model on the Word2Vec model.

A more detailed inspection of the graph allows us to identify patterns in the groups. First, we can say that the news from

TABLE IV

PRECISION RESULTS FOR CLASSIFICATION MODEL

Approach	Precision
Word2Vec	0.78
Word2Vec + K-Means	0.85

groups related to entertainment (groups 7, 8, and 2) are found at the opposite end of the news that talks about politics and corruption (group 1). Furthermore, if we make a more detailed inspection of the groups associated with sports (groups 7 and 8), we can identify patterns in the news they contain. In Figure 5, groups 6, 7, 8, and 2 are found. Groups 7 and 8 are quite similar, both covering sports in general, with football being the most predominant among them. Group 7 covers more football in general, being mostly and towards the center news about European or South American leagues. Group 8 exclusively covers news related in some way to the 2018 Russia World Cup, referencing different countries, including Peru. Group 6 covers national news, being largely news about Lima referring to different fields such as elections, activities, and some crimes. Group 2 covers news about social networks, commenting on publications of other Peruvian or international celebrities.

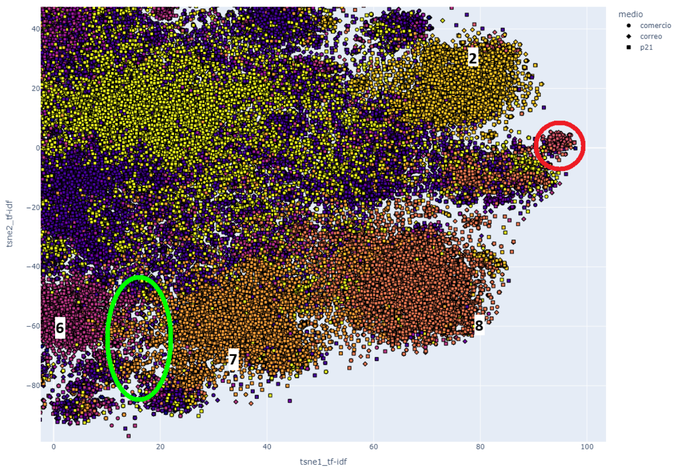

Fig. 5. Group associated with sports

The behavior we consider noteworthy is the interaction of these groups at their borders. At the border of groups 6 and 7 (enclosed in a green circle in Figure 5) are most of the articles covering Peruvian football, being a midpoint between national news and football news in general. As you move to the right, sports news starts to cover more international news. The other point to highlight is found between groups 8 and 2 (enclosed in a red circle in Figure 5). This group that serves as a bridge between sports (football) and social networks comprises mostly memes on social networks from different matches.

Moreover, it's worth noting that proximity of news doesn't always indicate similarity, as in several observations, the closest news doesn't necessarily cover the same topic. A constant in this aspect is the amount of noise present not only in the graph in general but also within the groups themselves.

As with TF-IDF, an inspection of the internal structure of each group was made. We observed that the behavior at the boundaries between groups is maintained in this model, in a cleaner form. The way this group interacts with the competitive sports group is notable, as the latter acts as an extension of the original group. Figure 6 shows groups 3 and 10 (associated with sports and competitive sports). In the focused fragment is specifically the part of the group that talks about football. It's worth noting that this overlap of groups is present in each area where sports are found. In the case of Figure 6, group 10 is separated into mini-groups that represent each competition of the closest sport. In the case of football, these would be leagues and tournaments at national and international levels. In order of appearance, from left to right, the competitions described by each group are: Liga 1 Perú Movistar, Copa Libertadores, Liga MX, 2018 Russia World Cup, and La Liga.

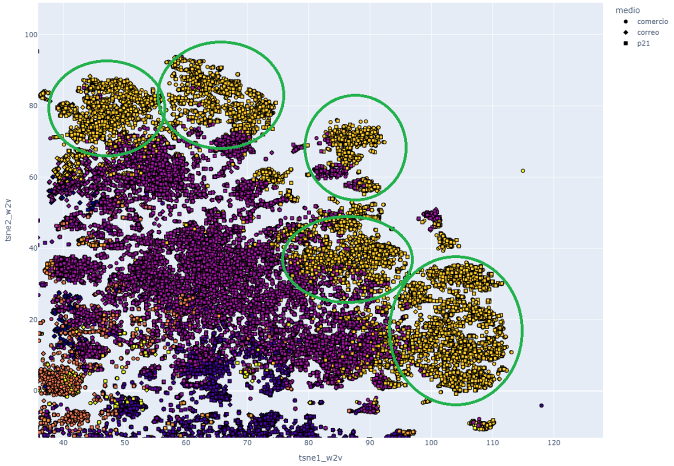

Fig. 6. Group associated with sports in the Word2Vec model

This behavior is replicated in each of the groups. In group 1 (corruption), they are separated into cases or characters, or in group 9 (entertainment) by franchises or types of entertainment, such as video games or movies. One last thing to highlight about this representation is the way there are few intruders in each of the clusters.

Choosing the Multilayer Perceptron model, based on the bibliography [11], demonstrated performance consistent with case expectations. The best-performing model was the one that used embedding through K-Means for classification. This technique has shown better performance in creating classification models, as observed in the previous section.

However, some challenges were encountered. The elbow test for determining the optimal number of clusters indicated a continuously decreasing mean error, suggesting a high number of topics in the dataset. This aligns with the expectation given the large volume and diversity of articles over two years. Therefore, we opted for a lower number of clusters for initial exploration, focusing on a more general classification.

Another challenge was the presence of empty bodies in the articles, which were mostly associated with news primarily containing videos or photos. These constituted a very small part of the total dataset and were removed to maintain the dataset's integrity.

Finally, the comparison of MLP performance using different vectorization techniques highlighted the strengths of combining Word2Vec with K-Means, especially in minimizing false positives, which is crucial for reliable pair predictions in the context of bias analysis.

The use of Word2Vec in combination with K-Means has proven to be a better strategy for classification than Word2Vec. This is because Word2Vec is able to capture semantic relationships between words by converting them into high-dimensional vectors, which facilitates the identification of patterns and similarities in texts. By applying K-Means, these relationships are clustered so that texts with similar content are closer together in the vector space, thus improving the quality of the MLP model predictions.

The MLP model with Word2Vec showed an precision of 0.78 average, which is significant given the inherent challenge of classifying large and varied texts such as news articles. This value indicates that the model has a high ability to correctly identify whether two news items are about the same event, minimizing the false positive rate. This precision is crucial in the context of media bias analysis, since misclassifying two news stories as similar when they are not could lead to incorrect conclusions about event coverage.

On the other hand, combining Word2Vec with K-Means raised the precision to 0.85, which is a considerable increase. This increase reflects the effectiveness of K-Means in clustering word vectors so that patterns and themes within articles are clearer and more distinguishable. This additional clustering technique strengthens the ability of the MLP model to perform more accurate classifications by providing a more ordered and coherent structure in the input data.

However, it is necessary to check some of the results. It would be necessary to review the number K of clusters, given that the ideal number ends up being 3, since as it increases, the precision of the MLP model in test and train ends up being very high, which is an indicator of overfitting. In the other hand, the labeling of the data could be improved with more data and with more topics. A considerable quantity of them are from specific topics such as soccer and politics. It could be done with more data.

VI. Conclusion

In conclusion, this study demonstrates that spatial representation and data exploration are vital for understanding and processing large text datasets, such as news articles. The Word2Vec technique proved superior to TF-IDF in terms of coherent and precise clustering. Additionally, the combination of Word2Vec with K-Means significantly improved the performance of the MLP classification model, ensuring reliable and accurate pair predictions.

The development of the MLP model proved to be quite accurate for the news dataset presented in the train and test trials. In this field, we can specify that Word2Vec tends more towards overfitting than the Word2Vec and K-Means model, while the K-Means model is more precise and, at first glance, more general. However, this can be affected by the technique used for data labeling, as well as the parameters chosen for Word2Vec.

The findings also underscore the importance of optimizing preprocessing steps to handle diverse data formats effectively, thereby improving the overall analysis and insights derived from the dataset. This could include more sophisticated techniques for handling missing data, noise reduction in text, and adaptive preprocessing based on the specific characteristics of different news sources.

This study contributes to the field by providing an approach to the large text classification problem through Word2Vec and Word2Vec+K-Means. The methodology presented here can be adapted and expanded for various natural language processing tasks beyond news article classification, such as sentiment analysis, content recommendation systems, or trend detection in social media.

By continuing to refine these techniques and exploring new methodologies, researchers can further improve the precision and efficiency of large-scale text analysis, leading to more robust and insightful understanding of textual data across various domains.

References

[1] "Prensa digital: ingresos a nivel mundial 2017-2028 — Statista". Statista. Accessed May 15, 2024. [Online]. Available: https://es.statista.com/estadisticas/1425488/ingresos-de-la-prensa-digital-en-todo-el-mundo/

[2] Jiang, J. Y., Zhang, M., Li, C., Bendersky, M., Golbandi, N., & Najork, M, "Semantic text matching for long-form documents". The world wide web conference, pp. 795-806, May 2019.

[3] Khder, M. A. "Web scraping or web crawling: State of art, techniques, approaches and application". International Journal of Advances in Soft Computing & Its Applications, pp. 144-168. 2021

[4] B. Liu, T. Zhang, D. Niu, J. Lin, K. Lai, and Y. Xu, "Matching long text documents via graph convolutional networks," arXiv preprint arXiv:1802.07459, 2018.

[5] Koloski, B., Pollak, S., Škrlj, B., & Martinc, M., "Extending neural keyword extraction with TF-IDF tagset matching". In Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, pp. 22-29. 2021

[6] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", 2018. doi:10.48550/arXiv.1810.04805.

[7] S. Aftan and H. Shah, "A Survey on BERT and Its Applications," 2023 20th Learning and Technology Conference (L&T), Jeddah, Saudi Arabia, 2023, pp. 161-166, doi: 10.1109/LT58159.2023.10092289.

[8] Zhang, Q., Li, J., Jia, Q., Wang, C., Zhu, J., Wang, Z., & He, X., "UNBERT: User-News Matching BERT for News Recommendation". In IJCAI, pp. 3356-3362, August 2021

[9] Mozer, R., Miratrix, L., Kaufman, A. R., & Anastasopoulos, L. J., "Matching with text data: An experimental evaluation of methods for matching documents and of measuring match quality", Political Analysis, pp. 445-468. 2020.

[10] Bafna, P., Pramod, D., & Vaidya, A., ”Document clustering: TF-IDF approach”. In 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT),pp. 61-66. March 2016.

[11] Sen, M. U., Erdinc, H. Y., Yavuzalp, B., & Ganiz, M. C. ”Combining lexical and semantic similarity methods for news article matching”. In Data Science–Analytics and Applications: Proceedings of the 2nd International Data Science Conference–iDSC2019, pp. 29-35. 2019

[12] S. Bird, E. Loper, and E. Klein, “Natural language processing with python oreilly media inc,” 2009.

[13] Plisson, J., Lavrac, N., & Mladenic, D., ”A rule based approach to word lemmatization”. In Proceedings of IS, pp. 83-86. October 2004.

[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Advances in neural information processing systems, 2013, pp. 3111–3119.

[15] L. van der Maaten and G. Hinton, ”Visualizing data using t-SNE,” Journal of Machine Learning Research, vol. 9, no. 11, 2008.

[16] G. Hamerly and C. Elkan, ”Learning the k in k-means,” in Advances in Neural Information Processing Systems, vol. 16, 2003.

[17] T. Windeatt, ”Accuracy/diversity and ensemble MLP classifier design,” IEEE Transactions on Neural Networks, vol. 17, no. 5, pp. 1194-1211, 2006.