Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 139-142 recibido 25-03-2020 revisado 30-05-2020 aceptado 30-05-2020

UMUCorpusClassifier: Compilation and evaluation of linguistic corpus for Natural Language Processing tasks

UMUCorpusClassifier: Recolección y evaluación de corpus lingüísticos para tareas de Procesamiento del Lenguaje Natural

José Antonio García-Díaz[1], **Ángela Almela**[2],
Gema Alcaraz-Mármol[3], **Rafael Valencia-García**[1]
[1]Facultad de Informática, Universidad de Murcia
[2]Facultad de Letras, Universidad de Murcia
[3]Departamento de Filología Moderna, Universidad de Castilla-La Mancha
{joseantonio.garcia8, angelalm, valencia}@um.es
gema.alcaraz@uclm.es

Abstract: The development of an annotated corpus is a very time-consuming task. Although some researchers have proposed the automatic annotation of a corpus based on ad-hoc heuristics, valid hypotheses cannot always be made. Even when the annotation process is performed by human annotators, the quality of the corpus is heavily influenced by disagreements between annotators or with themselves. Therefore, the lack of supervision of the annotation process can lead to poor quality corpus. In this work, we propose a demonstration of UMUCorpusClassifier, a NLP tool for aid researches for compiling corpus as well as coordinating and supervising the annotation process. This tool eases the daily supervision process and permits to detect deviations and inconsistencies during early stages of the annotation process.
Keywords: Corpus compilation, Document classification

Resumen: La construcción de un corpus anotado es una tarea que consume mucho tiempo. Aunque algunos investigadores han propuesto la anotación automática basada en heurísticas, éstas no siempre son posibles. Además, incluso cuando la anotación es realizada por personas puede haber discrepancias entre los mismos anotadores o de un anotador consigo mismo que influyen en la calidad del corpus. Por tanto, la falta de supervisión sobre el proceso de anotación puede llevar a corpus con baja calidad. En este trabajo, proponemos una demostración de UMUCorpusClassifier, una herramienta PLN para ayudar a los investigadores a compilar corpus y también a coordinar y supervisar el proceso de anotación. Esta herramienta facilita la monitorización diaria y permite detectar inconsistencias durante etapas tempranas del proceso de anotación.
Palabras clave: Compilación de corpus, Clasificación de documentos

1 Introduction

Supervised learning is the machine learning task which consists of building a model capable of predicting output for a specific problem based on prior observation of a previously labeled data set (Singh, Thakur, and Sharma, 2016). Supervised learning has applications for solving document classification tasks, which consist of matching a set of documents with a set of predefined labels. The main idea behind this is that supervised learning models can infer new knowledge by establishing associations between the examples provided and the expected tags. However, supervised learning requires a sufficient number of labeled examples that model the problem domain and, at the same time, the number of examples should be enough to cluster the examples in two subsets, one for model learning, and another for evaluating

ISSN 1135-5948. DOI 10.26342/2020-65-22

its accuracy based on samples that are not seen during the training stage.

The development of an annotated corpus is a very time-consuming process. To facilitate this task, some researchers have used distant supervision as a method of getting automatically annotated data (Go, Bhayani, and Huang, 2009). Distant supervision consists in the automatic tagging of the whole dataset based on certain assumptions or heuristics; however, valid hypotheses cannot always be made. Furthermore, automatic annotated data that have not thoroughly been reviewed can lead to noise in the data, which introduces information that do not follow the relationship with the rest of the elements of the dataset. These noisy documents require even greater volume of data to minimize its effects.

Furthermore, even though the annotation process occurs manually, the quality of the corpus cannot be guaranteed. In this sense, Mozetič, Igor et al. (Mozetič, Grčar, and Smailović, 2016) conducted an experiment to measure the quantity and quality of tagged Twitter datasets to evaluate the impact of measuring accuracy in sentiment analysis-based classification experiments, and they found that not all corpus annotated manually had a strong degree of agreement among the annotators, which indicated poor-quality. To solve these drawbacks, we present UMUCorpusClassifier, a tool that assists researchers in compiling text corpus, and helps to coordinate the annotation process.

The remainder of this paper is organized as follows: Section 2 describes our proposal, and Section 3 contains information regarding studies based on corpus compiled with this tool, as well as the description of future lines of action.

2 System architecture

UMUCorpusClassifier is designed to work with the social network Twitter, which is a widely used network for compiling corpus in various branches of Natural Language Processing (NLP) (Pak and Paroubek, 2010). Twitter provides a developer API through which you can query the social network. The free version of the Twitter API has some restrictions regarding the number of calls that can be made in a time interval. Therefore, to create an account in the platform, users are required to link a Twitter API key with their account. These credentials are generated through the Twitter API website. Once users have created and validated their account, they can create a corpus by entering a search string and, optionally, a geolocation. The search string has the same flexibility and limitations as the Twitter API; that is, it is allowed to nest terms, do exact searches, or search for specific user accounts. A list of these operators can be found here[1]. In addition, users can specify that only user-initiated messages are obtained by filtering messages that are responses to other users' tweets. Once the search query is specified, the corpus are compiled automatically. From time to time, a scheduled cron job process queries new tweets based on keywords. UMUCorpusClassifier works effectively with timelines, making use of *since_id* parameters to get new tweets and avoid duplicate results. However, this feature can be omitted in cases where the user deliberately changes the search string.

Identifying duplicate tweets is a complex task. Although Twitter provides a mechanism called *retweet* that allows the content of messages written by other users to be disseminated and the identification of these messages is trivial, many users in the social network use copy and paste mechanisms so it is possible to find duplicate or virtually the same tweets. Also, hyperlinks in tweets are encoded differently with each new tweet due to Twitter's own hyperlink shortening mechanism. For this reason, we have made the decision to replace URLs with a fixed token, which makes it easier to identify certain tweets. In addition, we have added a mechanism to calculate the similarity of the texts. Being an experimental technology, tweets are not removed, but administrators are given the opportunity to combine the responses of the tweets at their discretion.

The next step is to assign each corpus a set of independent labels. They can be made using a set of predefined labels, such as *out-of-domain*, *positive*, *negative*, *neutral*, *do-not-know-do-not-answer* or define a new set of tags for the corpus. Each label is identified with a color and a name.

The corpus labeling process can be carried out manually by the same user or allow access to the platform to a set of annotators and to supervise their work. Documents are

[1] https://developer.twitter.com/en/docs/tweets/rules-and-filtering/overview/standard-operators

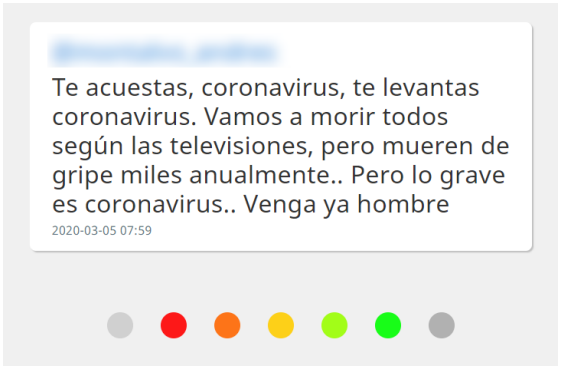

Figure 1: Screenshot of an annotator's view

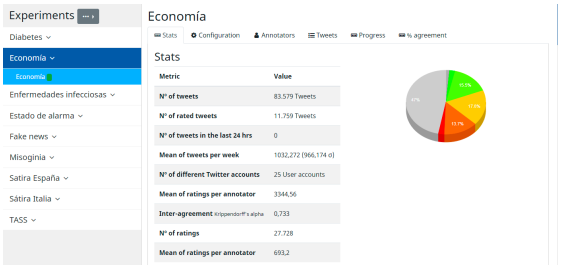

Figure 2: Screenshot of a researcher's view

annotated individually. When an annotator enters the web application, a tweet appears randomly from tweets that they had not previously classified. Figure 1 shows the view of an annotator at the platform.

To facilitate the labeling monitoring process, UMUCorpusClassifier provides a set of metrics and charts, such as the evolution of the annotations made by time, to evaluate that the work is being done constantly; the total rankings by tag; or (3) the average of annotations as well as their standard deviation. Figure 2 shows the researcher's view of the platform.

For each annotator, the degree of *self-agreement* is calculated, which measures how the same annotator classifies semantically similar documents. To cluster similar documents, we obtain the *sentence-embeddings* from FastText in its Spanish version (Grave et al., 2018) and then we have calculated the cosine distance among those vectors. Tweets that exceed a certain threshold distance are discarded.

For each tweet, one can see the degree of *inter-agreement* that measures how the same tweet is classified by different annotators by using the Krippendorff's alpha coefficient (Krippendorff, 2018). Moreover, inter-agreement has been proposed as upper bound

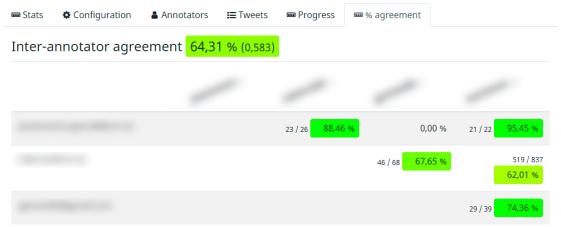

Figure 3: Screenshot of an annotator's view

to estimate the expected *accuracy* when finding a classifier to solve the classification problem (Mozetič, Grčar, and Smailović, 2016). Figure 3 shows the overall inter-agreement of the corpus, as well as the inter-agreement for each pair of annotators.

Once the corpus has been compiled and annotated, it can be exported to text formats. The export process is flexible and allows you to choose the number of classes to export. Combining classes is also allowed. This is useful, for example, when a classification has been made on a scale of very negative, negative, neutral, positive, and very positive type values, but one may want to combine the results to return the corpus grouped into positive, neutral, and negative. Corpora can be exported balanced, that is to say, the system automatically searches for the class with the largest number of instances and cuts instances from the other classes. These deleted instances are agreed by consensus. Finally, it is possible to export only the Twitter IDs in order to share them with the community as recommended by Twitter's privacy policies[2].

Furthermore, the number of instances to export can be selected and set. One of the advantages of this approach is that corpus can be exported by consensus: since the same tweet can be classified by different annotators, the number of tweets to export can be limited and retrieve those tweets that have achieved strong consensus among annotators. Thus, subsets of the corpus comprising the documents with common agreement can be retrieved, and the rest of the documents can be analyzed. Furthermore, the software is easily extensible. In this respect, it is relatively easy to include new strategies to export the data or to improve the platform to include new data sources other than Twitter.

[2]https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases

3 Further work

In this study, we have presented UMUCorpusClassifier, a NLP tool that assists in the compilation and annotation of linguistic corpus. So far, we have used this application on several domains. Specifically, we have compiled tweets about different types of diseases to carry out infodemiology studies that involve measuring the population's perception of infectious diseases (Apolinardo-Arzube et al., 2019; García-Díaz, Cánovas-García, and Valencia-García, 2020; Medina-Moreira et al., 2018), diabetes (Medina-Moreira et al., 2019), and figurative language such as satire (Salas-Zárate et al., 2017).

In the current version of the platform it is only possible to assign a label to a document. We are working to enable the *multi-label* classification. Another line of research is the addition of a contextual feature extraction module, enabling the analysis of groups of Twitter accounts from which tweets are extracted. These features may include information on the time of publication, number of followers, etc. Lastly, with regard to semantic similarity, we are currently analyzing ways to distance the most different tweets from each other, so that we can export the tweets with strongest consensus and the most distant ones.

Acknowledgments

This demonstration has been supported by the Spanish National Research Agency (AEI) and the European Regional Development Fund (FEDER/ERDF) through projects KBS4FIA (TIN2016-76323-R) and LaTe4PSP (PID2019-107652RB-I00). In addition, José Antonio García-Díaz has been supported by Banco Santander and University of Murcia through the Doctorado industrial programme.

References

Apolinardo-Arzube, O., J. A. García-Díaz, J. Medina-Moreira, H. Luna-Aveiga, and R. Valencia-García. 2019. Evaluating information-retrieval models and machine-learning classifiers for measuring the social perception towards infectious diseases. *Applied Sciences*, 9(14):2858.

García-Díaz, J. A., M. Cánovas-García, and R. Valencia-García. 2020. Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in latin america. *Future Generation Computer Systems*, 112:614–657.

Go, A., R. Bhayani, and L. Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Grave, E., P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. 2018. Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

Krippendorff, K. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Medina-Moreira, J., J. A. García-Díaz, O. Apolinardo-Arzube, H. Luna-Aveiga, and R. Valencia-García. 2019. Mining twitter for measuring social perception towards diabetes and obesity in central america. In *International Conference on Technologies and Innovation*, pages 81–94. Springer.

Medina-Moreira, J., J. O. Salavarria-Melo, K. Lagos-Ortiz, H. Luna-Aveiga, and R. Valencia-García. 2018. Opinion mining for measuring the social perception of infectious diseases. an infodemiology approach. In *Proceedings of the Technologies and Innovation: 4th International Conference, CITI*, page 229. Springer.

Mozetič, I., M. Grčar, and J. Smailović. 2016. Multilingual twitter sentiment classification: The role of human annotators. *PloS one*, 11(5).

Pak, A. and P. Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326.

Salas-Zárate, M. d. P., M. A. Paredes-Valverde, M. Á. Rodríguez-García, R. Valencia-García, and G. Alor-Hernández. 2017. Automatic detection of satire in twitter: A psycholinguistic-based approach. *Knowl. Based Syst.*, 128:20–33.

Singh, A., N. Thakur, and A. Sharma. 2016. A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1310–1315. Ieee.