Webology (ISSN: 1735-188X)
Volume 19, Number 6, 2022

Named-Entity Recognition Detector And Academic News Classification

Sir-Alexci Suarez Castrillon[1], Albert Miyer Suarez Castrillon[2], Jose Julian Cadena Morales[3]

[1] Faculty of Engineering, GRUCITE Research Group, University Francisco de Paula Santander Ocaña, Colombia.

[2] Faculty of Engineering and Architecture, GIMUP Research Group. Universidad de Pamplona, Colombia.

[3]Faculty of Education, Arts and Humanities. GIFEAH Research Group. University Francisco de Paula Santander Ocaña, Colombia.

ABSTRACT

A research is presented to classify the academic news according to 12 academic entities, which represent the programs of 4 faculties: Engineering, Arts and Humanities, Agricultural and Environmental Sciences, Administrative and Economic Sciences. And 12 academic/administrative entities. The purpose is that the news and documents published in the news archive are stored and classified according to the academic entity, which are evaluated by the peer reviewers of the Ministry of Education at the time of requesting relevant information and that has been socialized by each unit. A supervised classification is performed and the results show that the recognition of academic entities allows an adequate classification of all the news by department; however, the news of the academic/administrative processes are not completely classified because it does not identify the corresponding entity, but taking into account that the information requested is of the entities that are only academic, the objective set with the classification is achieved.

Keywords: entity naming; academic departments; supervised classification; natural language processing.

1. INTRODUCTION

The detection of entities in the documents is very important because they give us an idea of the importance they have in the document, and can help to classify the documents according to the needs (ICC, 2021), within the entities can be distinguished if they are Organization, Person, Location and Event, but they can also be determined by the context of the subject or according to the company or institution, and can be applied to education, security, medicine, among others.

This type of recognition is called Named Entity Recognition (NER), and the same entity can have a different meaning in the document according to the context (Baciero Fernandez, 2020). For example, it is used in legal texts in reference to laws, decrees, norms in legislative texts in peninsular Spanish (Samy, 2021). Although the easiest way to recognize entities is through a supervised classification, several works have allowed to know that through an unsupervised classification it is possible to improve the recognition in the English language, that is why some works try to transpolate the same methodology to the Spanish language, in order to know if it improves the classification results, demonstrating that it can be feasible when using a Conditional Random Field classifier (Copara Zea, 2017). If the entities are combined with the opinions about them, it is possible to know when a tourist site has a good reputation (Peres Estevan, 2015). One of the problems that can solve the recognition of entities, is the classification of documents through their information, which is dispersed in the network or framed within a general news system.

Continuing with the above, it is intended to provide a solution to the departments of an educational institution such as the Universidad Francisco de Paula Santander Ocaña where the news of each of its departments is published in a news archive without a proper classification, therefore, at the time of collecting the information that may be requested for a qualified registration, there is no knowledge of who the public is. It is very important for the departments to compile this information is sent to the news system since it must be stored for review and approval by the academic peers that supervise each program, a total of 12 departments must send this information, that is why if all the information goes out in a history without classification, the information may not be available neither by department nor by relevance. The objective is to be able to classify the information according to the academic entity that originates the news and to be present for its evaluation according to the requirements of the peer reviewers sent by the Ministry of Education.

2. METHODOLOGY

The information is classified by means of the standard entities, which are: organization, Person, Location, Event, Consumer Good and Work of Art (Figure 1), which the Salience value is analyzed, and it is taken into account if the supervised entity appears, in this case they are university entities based on academic and academic/administrative, having 12 academic entities and the rest of news are classified as academic/administrative (Table 1). The information is compiled from news published in the month of September 2022 from the news archive published on the UFPSO website.

Figure 2 shows the classification process, based on a supervised classification, where the entities to be recognized are already predetermined, as shown in Table 1.

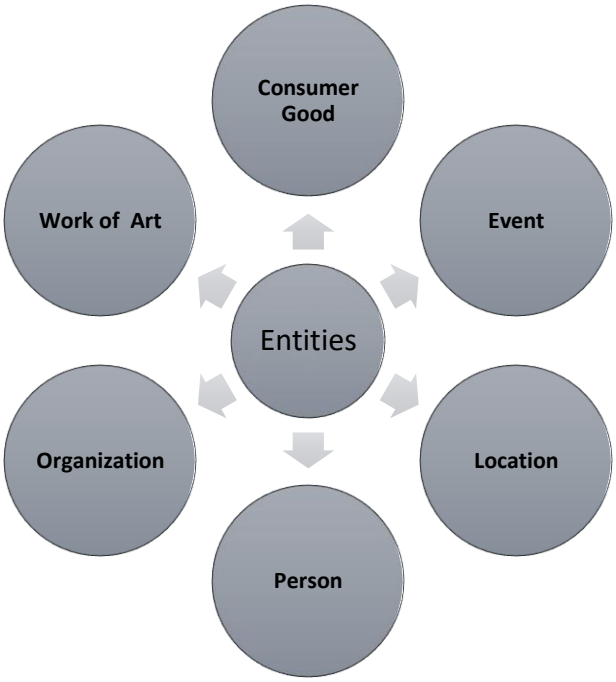

Figure 1. Standard entities.

Table 1. Academic Entities - Academic/Administrative.

ACADEMICS	**ACADEMIC/ADMINISTRATIVE**
Business Administration	Systems Division
Social Communication	Self-evaluation
Public Accounting	Internal Control
Law	Internships
Environmental Engineering	Experimental Farm
Civil Engineering	Personnel Division
Systems Engineering	Research Process
Mechanical Engineering	Extension Process
Commercial and Financial Management Technology	Planning Office
Animal Husbandry	Academic Subdirection
Academic and administrative processes	International Relations Office

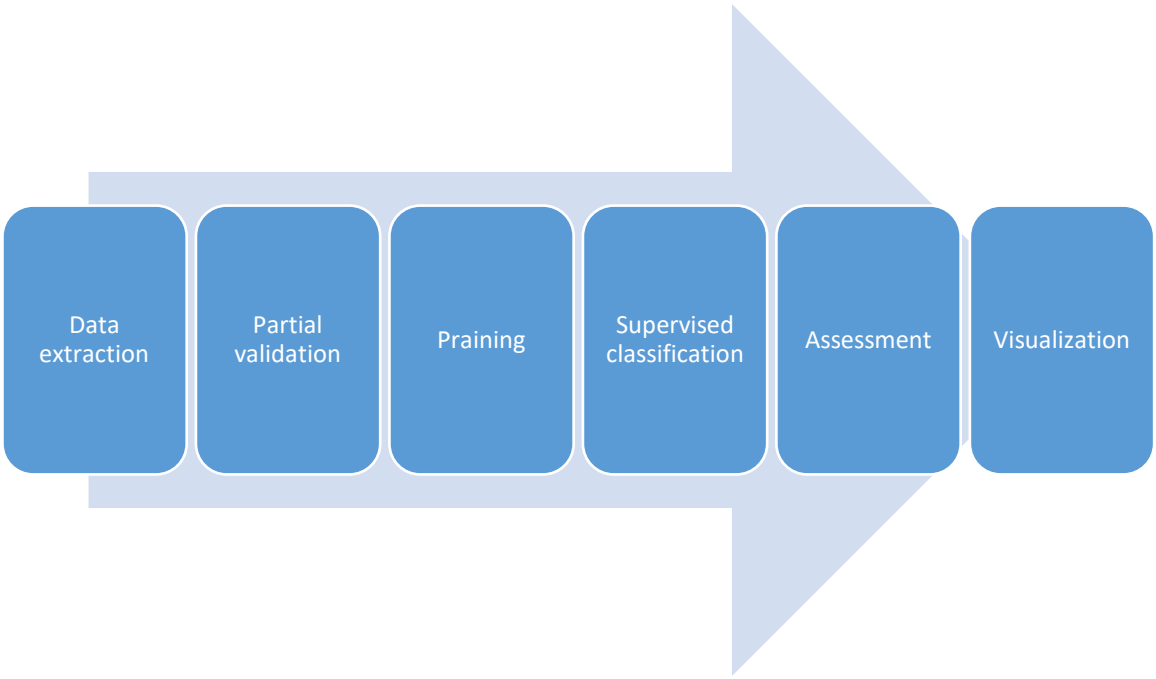

Figure 2. Classification process.

3. RESULTS

In table 2, the news can be classified by academic program in the Zootechnical program, it also tells us the location of the event and the people it is directed to, however the level of importance or visibility is high for teachers, and it is not highlighted to whom it is directed to which are the children, where the visibility is only 0. 02, and where entities such as teachers and university maintain a very high Salience, which does not reflect the mission of the event, which should be directed to the community of the whole city and more importantly to the children, in Figure 3, the importance value of each entity can be observed.

Table 2. News from the Zootechnics program about the Baby Zoo.

News title: Baby Zoo was successfully carried out by the Zootechnics program (UFPSOa, 2022)					
#	Tokens	Entities	#	Tokens	Entities
1	Teachers	Persons	2	University	Persons
3	Farm	Location	4	animals	Other
5	space	Other	6	Ocaña	Persons
7	Event	Event	8	Zootechnics Program	Other
9	Part	Persons	10	Children	Persons

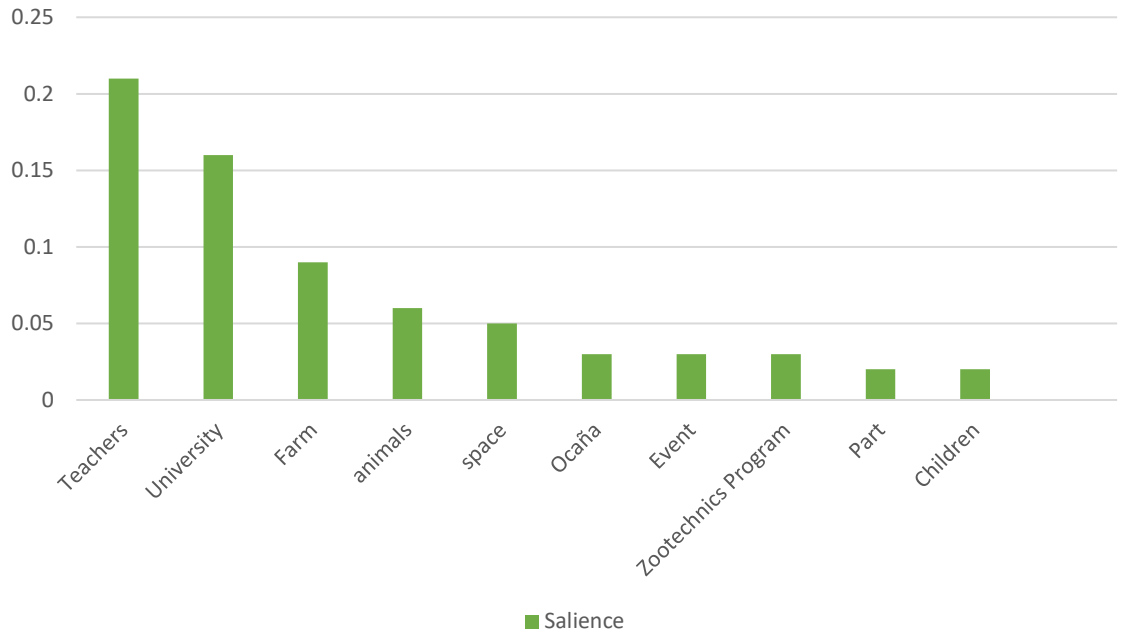

Figure 3. Salience for Baby Zoo news.

In table 3, the importance is high for the environmental engineering career, which allows classifying the information by program, however the news is its entities and importance does not highlight the event to be carried out, which is international mobility, although the university where the exchange will take place and the country appear.

Tabla 3. Placement according to Saliencia score, Environmental Engineering news.

News title: Environmental Engineering student performs international mobility (UFPSOb, 2022)					
#	Tokens	Entities	#	Tokens	Entities
1	Program	Other	2	University	Location
3	Environmental Engineering	Other	4	Institution	Organization
5	process	Other	6	Diego Alejandro	Persons
7	presence	Other	8	student	Persons
9	Universidad de Jaén	Organization	10	España	Location

In Figure 4, the value of salience is already very good to classify it in the academic entity of Environmental Engineering, with which the supervised classification works correctly, it is also observed the high value given to the program, with which it is easy to locate the news in the classification.

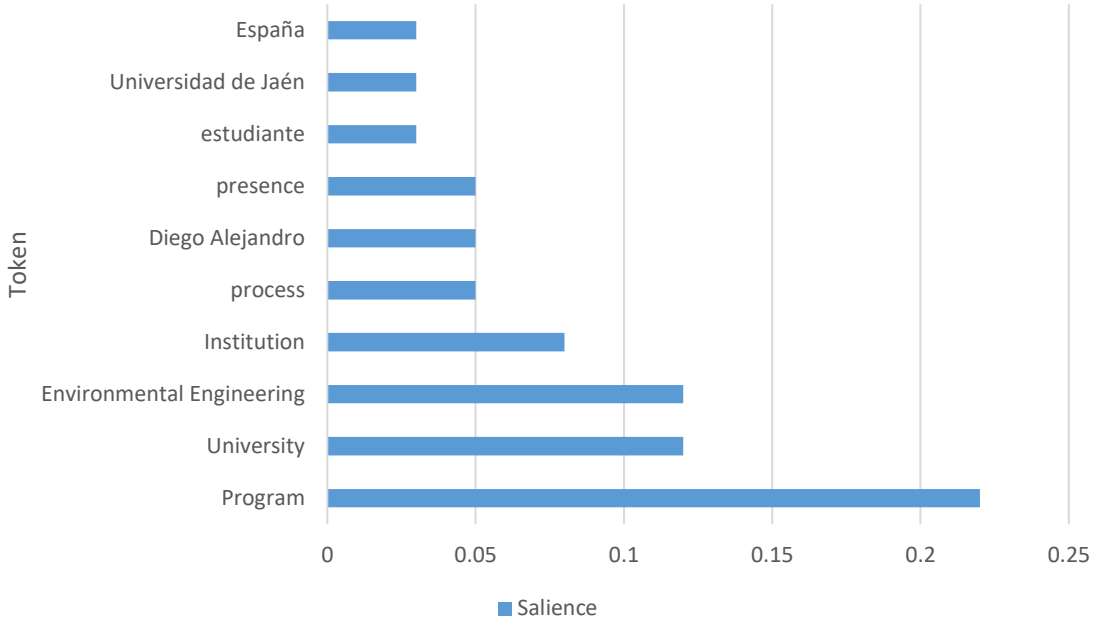

Figure 4. Environmental Engineering News.

In Table 4, it can be deduced that the entity of the news manages to classify the Social Communication program as the one with the highest visibility and also to highlight art as an entity, on which part of the program is based, although it is not presented as such an entity that tells us about the event, so it would fail if we want to classify it in any event category.

Table 4. Placement according to Salience score, Social Communication news.

News title: Cine Foro 35mm, began cycle of screenings during the current semester (UFPSOc, 2022)					
#	Tokens	Entities	#	Tokens	Entities
1	Social Communication Syllabus	Other	2	The Dignity of the Nadies	Location
3	Part	Other	4	opening	Event
5	planned	Other	6	agreement	Other
7	35mm Forum Cinema	Location	8	film	Work of art
9	projection	Other	10	Adamn	Persons

In Figure 5, it can be determined that the salience value for the academic entity is 28%, and is located in the first place, with which the news is classified correctly, indicating that the news of the Social Communication program are giving the appropriate relevance to be socialized and published, also the entity on the event maintains high percentages of 22%, which can be classified according to the topic by event.

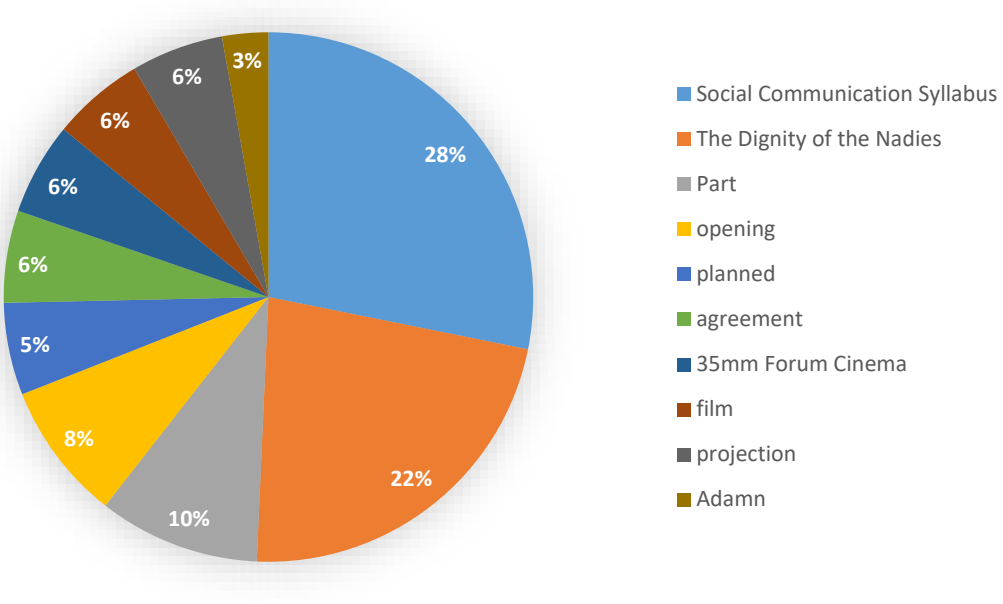

Figure 5. Salience value for Social Communication news.

The internal control process carried out an authoring event, however, as an entity it does not appear in the top 10 with the highest visibility and only highlights the place, although it is clear that it is an event, therefore it could not be classified concretely based on the importance of the internal control token, obtaining a rather poor result of 0 (Table 5).

Table 5. Location according to Salience score, Internal Control news.

News title: Internal Control advanced socialization of the audit program for the year 2022 (UFPSOd, 2022).							
#	Tokens	Entities	Visibilidad	#	Tokens	Entities	Visibilidad
1	audits	Event	0.23	2	Institutional Committee	Event	0.22
3	Program	Other	0.12	4	Auditorium	Location	0.04
5	Processes	Other	0.04	6	Socializatio n	Event	0.03
7	system	Other	0.03	8	Colombian technical standard	Other	0.03
9	installations	Other	0.02	1 0	Fabio Amaya	Persons	0.02

The extension process publishes the results of the call for projects for the year 2022-2023, and as an academic-administrative process it has a high visibility, occupying second place after the main objective which is the call for projects, therefore the news can be classified within the process of the Research and Extension Division (Table 6).

Table 6. Location according to Salience score, Extension news.

News title: Results of the Call for Extension Process (UFPSOe, 2022).							
#	Tokens	Entities	Visibilidad	#	Tokens	Entities	Visibilidad
1	projects	Other	0.50	2	extension	Location	0.07
3	contemplated	Other	0.03	4	convocation	Event	0.03
5	relationship	Other	0.03	6	calendar	Other	0.03
7	agreement	Location	0.03	8	UFPSO	Work of art	0.03
9	GIGMA	Other	0.02	10	process	Persons	0.02

Table 7 shows the values of entities according to the default classification where it is observed that the administrative news does not present a correct classification, but the news of academic entities are recognized in their totality.

Table 7. Salience values for news 1 to 5.

ENTITIES	**NEWS**				
	1	2	3	4	5
Business Administration	0	0	0	0	0
Social Communication	0	0	0.20	0	0
Public accounting	0	0	0	0	0
Law	0	0	0	0	0
Environmental Engineering	0	0.12	0	0	0
Civil Engineering	0	0	0	0	0
Systems Engineering	0	0	0	0	0
Mechanical Engineering	0	0	0	0	0
Commercial and Financial Management Technology	0	0	0	0	0
Animal Husbandry	0.03	0	0	0	0
Academic and administrative processes	0	0	0	-1	0.07

In Table 8, it can be seen that the news of the academic units can be classified in the correct units, but in the news 8 and 10 of the academic/administrative processes it is not clear to whom the information is addressed, it cannot be classified in any entity, so the wording confuses the reader.

Table 8. Salience values for news 6 to 10.

ENTITIES	**NEWS**				
	6	7	8	9	10
Business Administration	0	0	0	0	0

Social Communication	0	0	0	0	0
Public accounting	0	0	0	0	0
Law	0	0	0	0	0
Environmental Engineering	0.04	0	0	0	0
Civil Engineering	0	0	0	0	0
Systems Engineering	0	0	0	0.02	0
Mechanical Engineering	0	0	0	0	0
Commercial and Financial Management Technology	0	0	0	0	0
Animal Husbandry	0	0	0	0	0
Academic and administrative processes	0	0.22	-1	0	-1

The information of the news 11 to 15 is properly classified (Table 9). Thus, the classification by academic entity is 100% correct, however, for academic/administrative processes it may end up in an error, so according to the way of publication of the news, the system can only classify academic entities, and finally they are the main basis of all qualified registration, since the news are based on the socialization of these entities, and not on the academic/administrative ones.

Table 9. Salience values for news 11 to 15.

ENTITIES	**NEWS**				
	11	12	13	14	15
Business Administration	0.10	0	0	0	0
Social Communication	0	0	0	0	0
Public accounting	0	0	0	0.57	0
Law	0	0	0	0	0
Environmental Engineering	0	0	0	0	0
Civil Engineering	0	0	0	0	0
Systems Engineering	0	0	0	0	0
Mechanical Engineering	0	0	0	0	0.11
Commercial and Financial Management Technology	0	0	0	0	0
Animal Husbandry	0	0	0	0	0
Academic and administrative processes	0	0.14	0.03	0	0

4. CONCLUSIONS

The recognition of entities is a way to classify documents that do not have a defined classification, and can help in the organization of academic information in different departments, while the administrative information that is published does not present a clear

definition of the entities, and this can cause the information to dissipate and not reach the target audience.

REFERENCES

Baciero Fernandez, J. I. (2020). Elaboración de un Modelo de Reconocimiento de Entidades Nominales (NER) para su uso en aplicaciones de Procesamiento del Lenguaje Natural (NLP) [E.T.S.I. Industriales (UPM)]. {https://oa.upm.es/62858/

Copara Zea, J. L. (2017). Reconocimiento de entidades nombradas para el idioma español utilizando Conditional Random Fields con características no supervisadas. http://repositorio.concytec.gob.pe/handle/20.500.12390/1946

ICC. (2021). Procesamiento del Lenguaje Natural. Instituto de Ingeniería del Conocimiento. https://www.iic.uam.es/inteligencia-artificial/procesamiento-del-lenguaje-natural/

Peres Estevan, F. J. (2015). Análisis de la reputación de un destino turístico en las redes sociales. https://www.academia.edu/24121208/An%C3%A1lisis_de_la_reputaci%C3%B3n_de_un_destino_tur%C3%ADstico_en_las_redes_sociales

Samy, D. (2021). Reconocimiento y clasificación de entidades nombradas en textos legales en español. https://doi.org/10.26342/2021-67-9

UFPSOa, C. de M.-. (2022, septiembre 29). UFPSO - Baby Zoo fue realizado con éxito por el programa de Zootecnia. https://ufpso.edu.co/new/Baby-Zoo-fue-realizado-con-exito-por-el-programa-de-Zootecnia_4588

UFPSOb, C. de M.-. (2022, septiembre 28). UFPSO - Estudiante de Ingeniería Ambiental realiza movilidad internacional. https://ufpso.edu.co/new/Estudiante-de-Ingenieria-Ambiental-realiza-movilidad-internacional_4587

UFPSOc, C. de M.-. (2022, septiembre 28). UFPSO - Cine Foro 35mm, inició ciclo de proyecciones durante el presente semestre. https://ufpso.edu.co/new/Cine-Foro-35mm,-inicio-ciclo-de-proyecciones-durante-el-presente-semestre_4586

UFPSOd, C. de M.-. (2022, septiembre 28). UFPSO - Control Interno adelantó socialización del programa de auditorías vigencia 2022. https://ufpso.edu.co/new/Control-Interno-adelanto-socializacion-del-programa-de-auditorias-vigencia-2022_4585

UFPSOe, C. de M.-. (2022, septiembre 27). UFPSO - Resultados Convocatoria Proceso de Extensión. https://ufpso.edu.co/new/Resultados-Convocatoria-Proceso-de-Extension_4583