ALIADA: Artificial Intelligence-based language applications for the detection of aggressiveness in social networks

ALIADA: Aplicaciones del Lenguaje basadas en Inteligencia Artificial para la Detección de la Agresividad en Redes Sociales

José Alberto Mesa Murgado[1], Flor Miriam Plaza-del-Arco[1], Jaime Collado-Montañez[1], L. Alfonso Ureña-López[1] and M. Teresa Martín-Valdivia[1]

[1] *Departamento de Informática, CEATIC, Universidad de Jaén, España*

Abstract

In this paper, we present a Web Application Platform for the Detection of Aggressiveness in Social Media using Natural Language Processing and Machine Learning techniques, describing its architecture, the development technologies used and the different language models that have been integrated into the system. Finally, we conclude that the platform is a powerful tool to tackle real time aggressiveness on social media such as sexism or hate speech.

Keywords

Aggressiveness Detection, Web Application, Natural Language Processing, Machine Learning, Deep Learning

1. Introduction

The misuse of the Internet and specifically of social networks as a powerful tool for dialogue and participation, can lead to the creation, proliferation and dissemination of hate speech. According to the report on the evolution of hate crimes in Spain in 2020[1], Internet (45%) and social networks (22.8%) are the most used means for the commission of hate speech, with messages of ideology, racism/xenophobia, sexual orientation and gender identity showing the highest incidence. Threats, insults and public promotion/incitement to hatred, hostility, discrimination are computed as the most repeated criminal acts. Other communication channels where these acts are committed, but to a lesser extent, are telephony/communications (14.3%) and other sources of social communication (4.2%). The high incidence of these crimes on the Internet and social media shows the high need to combat them. Detecting this phenomenon can help to social media moderators to warn/block bullies and provide support to victims.

In the last years, offensive language research has emerged in the Natural Language Processing (NLP) area seeking to offer solutions to detect automatically this inappropriate behavior on the Web [1, 2]. The most recent and best-performing studies offer solutions based on neural networks for the detection of the different phenomena including misogyny and xenophobia [3], sexism [4], cyberbullying [5], aggression [6], or offensive language [7, 8]. Some researches have shown that sentiment and emotion analysis are important features to consider in the detection of these phenomena [9, 10, 11]. Although more and more studies are being conducted in this area, the integration of these automatic models to be used in real scenarios by any user is very scarce, especially in languages other than English, such as Spanish.

In this paper we present ALIADA, an artificial intelligence-based language application for the detection of aggressiveness[2] in social media. This application allows real-time monitoring of viral events on the social networks: Youtube and Twitter, integrating trained language models based on NLP solutions to identify aggressiveness on this content and visualizing the outcome to the user. In addition, to overcome the lack of language models available for offensive language research in Spanish, we have taken advantage of the majority of Spanish corpora that have been developed in this area to train different Machine Learning (ML) solutions for

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain

jmurgado@ujaen.es (J. A. M. Murgado); fmplaza@ujaen.es (F. M. Plaza-del-Arco); jcollado@ujaen.es (J. Collado-Montañez); laurena@ujaen.es (L. A. Ureña-López); maite@ujaen.es (M. T. Martín-Valdivia)

[1] https://bit.ly/3611hm9

[2] We use aggressiveness term to encompass different phenomena such as hate speech, sexism, misogyny, offense.

the detection of aggression in real-time data.

The rest of the paper is structured as follows: In Section 2 we provide a description of the tool and its architecture. Language models implemented are explained in Section 3. Finally, Section 4 presents conclusions and future work.

2. System Description

The ALIADA Web application consists of five internal modules that interact with each other to attend incoming requests and provide resources to relevant stakeholders (hereinafter, namely, users):

- **Data Storage Module**, based on ELK's Elasticsearch search engine it allows to index data under a non SQL approach.
- **Routing Module**, relies on the FastAPI framework to attend requests asynchronously using Python.
- **User Interface Module**, built using state-of-the-art web technologies such as HTML5, CSS3 (specifically, Bootstrap 5 as CSS framework) and Javascript.
- **Internal Logic Module**, implemented using Python manages data retrievals from social networking sites and the classification of incoming users requests.
- **Artificial Intelligence: Machine Learning Module**, built upon the Torch library for Python, allows to perform inferences in ML and Deep Learning models.

These modules are organized into Backend and Frontend, the former being responsible for routing and associated logic, and the latter of providing a graphical interface to interact with.

2.1. Backend

Encompasses the routing management and handling of incoming endpoint calls:

2.1.1. Stored Data and Storage Process

Information regarding users, their related personalization and data retrieval and classification requests, is stored in an Elasticsearch repository considering:

- The type of the submitted request: either data retrieval or classification.
- Social network used as source: Twitter or Youtube.
- The language model applied.

2.1.2. Data Retrieval and Extraction of New Data

Users can retrieve social data through requests, in which the social network used as source must be specified along with other search parameters: (1) who sent the post or (2) to whom it is targeted at, in which period of time it was published (3) or whether it includes an user provided keyword. Gathered data is anonymized before being stored in the Elasticsearch data warehouse in string format, structured as: (1) source, (2) corresponding source identifier, (3) parent source identifier, whether the publication is a response, (4) release date, and (5) associated textual content (tweet or comment).

Request's retrieved data can be downloaded in comma separated format (.csv) however, importing new data into a request is not allowed. At the same time, a request social data cannot be shared in other requests or by any other users distinct from their original requester who is allowed to run different ML classifying models against a same request in order to collect diverse statistics (e.g: in terms of sexism, offensiveness, hate speech, etc.).

2.1.3. User creation and management

Responses from the server require of authorized credentials that must be granted by an administrator, after requesting access through the contact form on the platform's homepage.

Users must be logged in to request and classify social data, this authorization is sent in each HTTP Request through Javascript Web Tokens (JWT) and serves two purposes: (1) security and (2) personalization.

2.1.4. Request Management

On the one hand, users' requests for data retrieval and classification are segmented into separated queues and serviced according to the date on which they were sent to the server along with a priority value that is reduced progressively as long as no new data is retrieved from the source, helping to determine when a certain topic is no longer relevant. On the other hand, the server traffic is handled asynchronously through FastAPI's uvicorn library which allows to run an ASGI Web server.

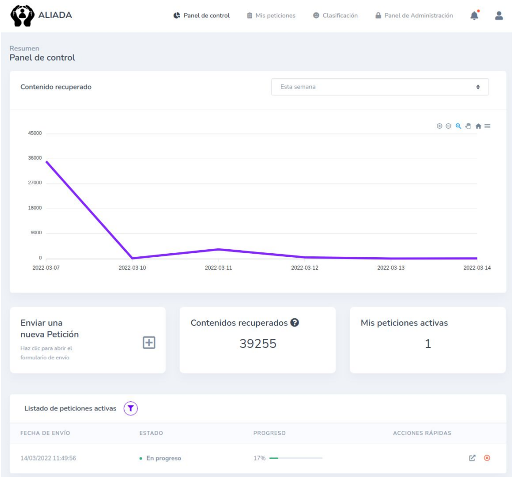
2.1.5. Data Classification and Procedure to Add New Models

Classification orders are associated to retrieval requests, they specify which ML model will be applied to the data and internally, they are ordered by the date in which they were sent to the server. Further on, the Pickle and Torch libraries are used to load the trained model architecture and state, as well as its associated vocabulary. Integrating new ML models into the server requires for the uploading of the trained model along with its corresponding word embeddings or bag-of-words structure and a categorical label dictionary to improve the comprehensibility of the model. A new function must be declared inside the Classifying module to load the model and use it against input data.

2.2. Frontend: User Interface

ALIADA provides a minimalistic web interface to make use of all of its features in a fast and intuitive way. Right after logging in from the main webpage, access to all the application's functions is provided: New data retrieval requests, statistics about the classification results, graphs of the total amount of downloaded posts, etc. In the following, these features are further described.

Dashboard. A dashboard (Figure 1) containing the current status of data retrieval and classification requests is displayed. Here, the client can see an ApexCharts' graph[3] that plots the total amount of data downloaded in a given time period, a list containing all active requests and a button to create a new one. Clicking on this button will pop up a form with all the information required to send a new data retrieval request as shown in Figure 2.

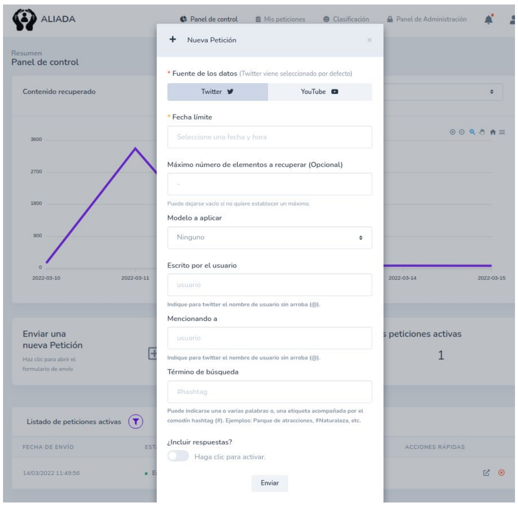


Contenidos recuperados 39255

Mis peticiones activas 1

FECHA DE ENVÍO	ESTADO	PROGRESO	ACCIONES RÁPIDAS
14/03/2022 11:49:56	● En progreso	17%	

Figure 1: Dashboard.



* Fuente de los datos (Twitter viene seleccionado por defecto)

* Fecha límite: Seleccione una fecha y hora

Máximo número de elementos a recuperar (Opcional)

Puede dejarse vacío si no quiere establecer un máximo.

Modelo a aplicar: Ninguno

Escrito por el usuario: usuario

Indique para twitter el nombre de usuario sin arroba (@).

Mencionando a: usuario

Indique para twitter el nombre de usuario sin arroba (@).

Término de búsqueda: #hashtag

Puede indicarse una o varias palabras o, una etiqueta acompañada por el comodín hashtag (#). Ejemplos: Parque de atracciones, #Naturaleza, etc.

¿Incluir respuestas? Haga clic para activar.

Figure 2: New request form.

My requests and classification panel. In order to have a more in-depth view of active and completed requests, two different sections are provided: my requests and classification panel. The former shows the current state (queued, in progress or completed) of each data retrieval request, while the latter shows the classification results in the form of graphs as seen in Figure 3. This section also shows all anonymized texts with their predicted labels, some information about the data retrieval and buttons to both download the full retrieved corpus as a .csv file and reuse the data to infer new labels with a different ML model.

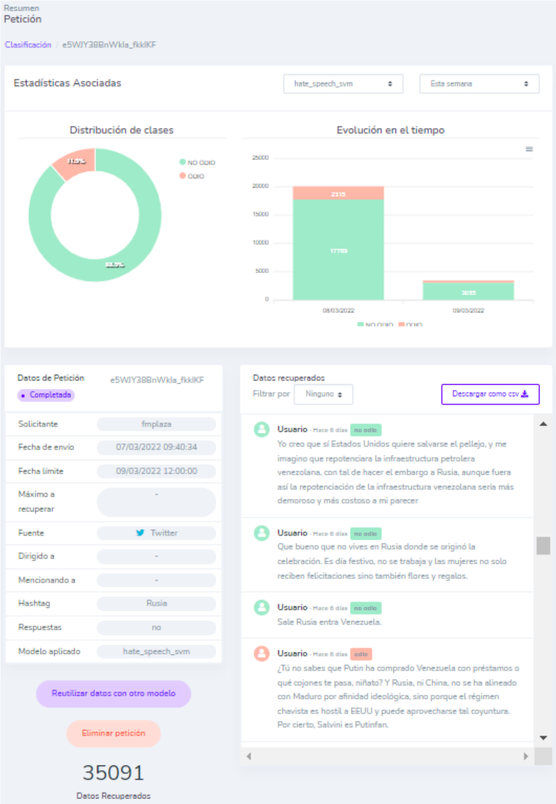
Administrator. Finally, only users with the administrator role have access to the administration panel. Here, an administrator can see the application's log history or the list of active requests in real-time. Users can also be created and deleted from this panel.

3. Language Models

The main objective of ALIADA is to monitor social media posts for the detection of aggressive content. Therefore, it is necessary to integrate different ML solutions to detect this behavior. Specifically, we have trained different models based on SVM for the detection of three phenomena: hate speech, sexism, and offensiveness.

In order to train these solutions, we have taken into account most of the available corpora generated for aggressiveness detection in Spanish including

[3] https://apexcharts.com/



Solicitante	fmplaza
Fecha de envio	07/03/2022 09:40:34
Fecha límite	09/03/2022 12:00:00
Máximo a recuperar	-
Fuente	Twitter
Dirigido a	-
Mencionando a	-
Hashtag	Rusia
Respuestas	no
Modelo aplicado	hate_speech_svm

Usuario - Hace 6 días no odio
Yo creo que sí Estados Unidos quiere salvarse el pellejo, y me imagino que repotenciara la infraestructura petrolera venezolana, con tal de hacer el embargo a Rusia, aunque fuera así la repotenciación de la infraestructura venezolana sería más demoroso y más costoso a mi parecer

Usuario - Hace 6 días no odio
Que bueno que no vives en Rusia donde se originó la celebración. Es día festivo, no se trabaja y las mujeres no solo reciben felicitaciones sino también flores y regalos.

Usuario - Hace 6 días no odio
Sale Rusia entra Venezuela.

Usuario - Hace 6 días odio
¿Tú no sabes que Putin ha comprado Venezuela con préstamos o qué cojones te pasa, niñato? Y Rusia, ni China, no se ha alineado con Maduro por afinidad ideológica, sino porque el régimen chavista es hostil a EEUU y puede aprovecharse tal coyuntura. Por cierto, Salvini es Putinfan.

Figure 3: Classification results.

HatEval [12], HaterNet [13], EXIST [14], NewsComTOX [15] and OffendES [7]. A total of four models are available in the platform: *hate_speech_svm* has been trained on HatEval, HaterNet and NewsComTOX datasets, *offendes_svm* has been trained on the large OffendES dataset, *sexism_svm* is trained on the EXIST dataset and finally *all_concepts_svm* combine all of the datasets.

4. Conclusions and Future Work

ALIADA is a powerful and useful tool to tackle aggressiveness in social networking sites in real-time, allowing for the detection of such attitudes in social publications through ML algorithms. In the near future, we would like to go further and, in addition to post classification, we will develop an explainability tool in order to understand what sections within each post makes it more aggressive than others through what is known as Named Entity Recognition (NER) techniques, and an emotion or performance tool to determine which attitude causes a greater effect in terms of its associated social reactions (namely, likes and retweets).

5. Acknowledgments

Acknowledgments

This work has been partially supported by Big Hug project (P20_00956, PAIDI 2020) and WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government, LIVING-LANG project (RTI2018-094653-B-C21) funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe, and the scholarship (FPI-PRE2019-089310) from the Ministry of Science, Innovation, and Universities of the Spanish Government.

References

[1] E. Fersini, P. Rosso, M. Anzovino, Overview of the Task on Automatic Misogyny Identification at IberEval 2018 (2018) 15.

[2] M. E. Aragón, M. Álvarez Carmona, H. J. Escalante, L. Villaseñor-Pineda, D. Moctezuma, Overview of MEX-A3T at IberLEF 2019: Authorship and aggressiveness analysis in Mexican Spanish tweets (2019) 17.

[3] F.-M. Plaza-del-Arco, M. D. Molina-González, L. A. Ureña López, M. T. Martín-Valdivia, Detecting Misogyny and Xenophobia in Spanish Tweets Using Language Technologies, ACM Trans. Internet Technol. 20 (2020). URL: https://doi.org/10.1145/3369869. doi:10.1145/3369869.

[4] F. M. Plaza-del-Arco, M. D. Molina-González, L. A. U. López, M. T. Martín-Valdivia, Sexism Identification in Social Networks using a Multi-Task Learning System, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021, volume 2943 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 491–499.

[5] F. Elsafoury, S. Katsigiannis, S. R. Wilson, N. Ramzan, Does BERT Pay Attention to Cyberbullying?, Association for Computing Machinery, New York, NY, USA, 2021, p. 1900–1904. URL: https://doi.org/10.1145/3404835.3463029.

[6] R. Kumar, A. K. Ojha, S. Malmasi,

M. Zampieri, Evaluating Aggression Identification in Social Media, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 1–5. URL: https://aclanthology.org/2020.trac-1.1.

[7] F. M. Plaza-del-Arco, A. Montejo-Ráez, L. A. Ureña-López, M.-T. Martín-Valdivia, OffendES: A New Corpus in Spanish for Offensive Language Research, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), INCOMA Ltd., Held Online, 2021, pp. 1096–1108. URL: https://aclanthology.org/2021.ranlp-1.123.

[8] F. M. Plaza-del-Arco, M. Casavantes, H. Escalante, M. T. Martin-Valdivia, A. Montejo-Ráez, M. Montes-y-Gómez, H. Jarquín-Vásquez, L. Villaseñor-Pineda, Overview of the MeOffendEs task on offensive text detection at IberLEF 2021, Procesamiento del Lenguaje Natural 67 (2021).

[9] S. Rajamanickam, P. Mishra, H. Yannakoudakis, E. Shutova, Joint Modelling of Emotion and Abusive Language Detection, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4270–4279. URL: https://aclanthology.org/2020.acl-main.394. doi:10.18653/v1/2020.acl-main.394.

[10] F. M. Plaza-del-Arco, M. D. Molina-González, L. A. Ureña-López, M. T. Martín-Valdivia, A Multi-Task Learning Approach to Hate Speech Detection Leveraging Sentiment Analysis, IEEE Access 9 (2021) 112478–112489.

[11] F. M. Plaza-del-Arco, S. Halat, S. Padó, R. Klinger, Multi-Task Learning with Sentiment, Emotion, and Target Detection to Recognize Hate Speech and Offensive Language, CoRR abs/2109.10255 (2021). URL: https://arxiv.org/abs/2109.10255. arXiv:2109.10255.

[12] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. Rangel, P. Rosso, M. Sanguinetti, SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), Association for Computational Linguistics, 2019.

[13] J. C. Pereira-Kohatsu, L. Quijano-Sánchez, F. Liberatore, M. Camacho-Collados, Detecting and Monitoring Hate Speech in Twitter, Sensors 19 (2019) 4654.

[14] F. Rodríguez-Sánchez, J. C. de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of EXIST 2021: sEXism Identification in Social neTworks, Procesamiento del Lenguaje Natural 67 (2021).

[15] M. Taulé, A. Ariza, M. Nofre, E. Amigó, P. Rosso, Overview of the DETOXIS Task at IberLEF-2021: DEtection of TOXicity in comments In Spanish, Procesamiento del Lenguaje Natural 67 (2021).