MODERATES: MODERAtion of ConTEnts in Social networks using Language Technologies

L. Alfonso Ureña-López[1], Eugenio Martínez Cámara[1], Salud María Jiménez-Zafra[1], Miguel Ángel García-Cumbreras[1], José M. Perea-Ortega[2], Arturo Montejo-Ráez[1], Manuel García-Vega[1], M.Dolores Molina González[1], Fernando Martínez-Santiago[1], Manuel Carlos Díaz-Galiano[1] and M.Teresa Martín-Valdivia[1]

[1]*Department of Computer Science, Advanced Studies Center in ICT (CEATIC), Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain*
[2]*Department of Computer and Telematic Systems, Universidad de Extremadura, Avda. Elvas s/n. 06006, Badajoz, Spain*

Abstract

The rise of digital platforms has led to an increase in social interactions, but it has also given rise to inappropriate behavior on the Web, including the spread of hate speech and offensive language. This poses a threat to freedom of expression, exposing users to denigrating content based on personal characteristics. Such communication can have harmful psychological effects, especially on vulnerable communities. Detecting and preventing hate speech has become a crucial area of research in Natural Language Processing (NLP) and Machine Learning (ML). This project aims to develop effective solutions using advanced ML and NLP techniques to assess the offensiveness and toxicity of Spanish text. A cloud service will automatically label and classify texts, providing an interface for detecting and substituting inappropriate language. The prototype aims to assist users in writing non-offensive content and serve as real-time monitoring and filtering tools for social media, contributing to a more inclusive and fair discourse.

Keywords

Natural Language Processing, NLP, human language technologies, language modeling, machine learning, offensive language and hate speech, sentiment and emotion analysis

1. Introduction

The significant increase in social interactions through digital platforms has leveraged the presence of inappropriate behavior on the Web, such as the propagation of hate speech or the use of offensive language among the users of these platforms. Freedom of expression in these media has exposed their users to publications that are sometimes used to denigrate, insult or hurt with foul or rude language based on gender, race, religion, ideology or other personal characteristics. Unfortunately, this type of communication can be very harmful and can cause negative psychological effects among users, especially among vulnerable communities such as children and teenagers, women, LGTBI, immigrants, religious and cultural communities.

Governments and online platforms have been addressing this problem for some years now, and measures are continuously being adopted in the form of laws and policies that contribute to fostering healthy coexistence in these media. For example, since 2013, the European Council has promoted the "No Hate Speech" movement [1] with the aim of mobilizing young people to combat hate speech and promote human rights on the Internet. In May 2016, the European Commission reached an agreement with Facebook, Microsoft, Twitter and YouTube to create a "Code of Conduct on combating illegal hate speech online"[2]. Based on this agreement, the "Protocol on combating illegal hate speech online" has been drafted[3].

Between 2018 and 2020, other platforms such as In-

SEPLN-CEDI-PD 2024: Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations, June 19-20, 2024, A Coruña, Spain.
laurenas@ujaen.es (L. A. Ureña-López); emcamara@ujaen.es (E. M. Cámara); sjzafra@ujaen.es (S. M. Jiménez-Zafra); magc@ujaen.es (M. García-Cumbreras); jmperea@unex.es (J. M. Perea-Ortega); amontejo@ujaen.es (A. Montejo-Ráez); mgarcia@ujaen.es (M. García-Vega); mdmolina@ujaen.es (M.Dolores M. González); dofer@ujaen.es (F. Martínez-Santiago); mcdiaz@ujaen.es (M. C. Díaz-Galiano); maite@ujaen.es (M.Teresa Martín-Valdivia)
0000-0001-7540-4059 (L. A. Ureña-López); 0000-0002-5279-8355 (E. M. Cámara); 0000-0003-3274-8825 (S. M. Jiménez-Zafra); 0000-0003-1867-9587 (M. García-Cumbreras); 0000-0002-7929-3963 (J. M. Perea-Ortega); 0000-0002-8643-2714 (A. Montejo-Ráez); 0000-0003-2850-4940 (M. García-Vega); 0000-0002-8348-7154 (M.Dolores M. González); 0000-0002-1480-1752 (F. Martínez-Santiago); 0000-0001-9298-1376 (M. C. Díaz-Galiano); 0000-0002-2874-0401 (M.Teresa Martín-Valdivia)

[1] https://www.coe.int/en/web/committee-on-combatting-hate-speech/home

[2] https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

[3] https://www.inclusion.gob.es/oberaxe/ficheros/ejes/discursoodio/PROTOCOLO_DISCURSO_ODIO.pdf

stagram, Snapchat, Dailymotion and TikTok joined this Code of Conduct. The 2019 report[4] highlighted that threats, insults and discrimination are counted as the most repeated criminal acts, with the Internet (54.9%) and social networks (17.2%) being the most used means to commit these actions. This situation has led the Spanish Parliament to approve in 2020 a law to prevent the spread of hate online. The 2020 report[5] maintains the trend, with the Internet (45%) and social networks (22.8%) as the media through which hate crimes are most assiduously disseminated. However, this problem does not only involve governments and online platforms, but also affects society in general, where the number of this type of crimes and aggressive behavior on the Internet has increased exponentially in recent years.

Thus, it seems clear that the problem of detecting inappropriate behavior in general, and hate speech in particular, on the Web has worsened in recent years, making it necessary to study, analyze and implement solutions in all areas, including the field of language technologies [1]. Analyzing this type of harmful content on the Web requires automatic systems capable of processing and analyzing human language[2]. For this reason, the detection and prevention of hate speech and offensive language has become one of the main research topics of Natural Language Processing (NLP). NLP is an important area of Artificial Intelligence that tries to understand and generate language in the same way as humans do using computational methods. In addition, the use of Machine Learning (ML) algorithms is allowing the development of powerful classification systems that, combined with advanced NLP techniques, help us to provide answers to many current social problems.

In addition, hope speech is a type of language that is able to relax a hostile environment and that helps, gives suggestions and inspires for good to a number of people when they are in times of illness, stress, loneliness or depression. Detecting it automatically, so that positive comments can be more widely disseminated, can have a very significant effect when it comes to combating sexual or racial discrimination or when we seek to foster less bellicose environments[3].

To train these systems, it is essential to generate and compile manually labeled linguistic resources as well as to adapt existing ones to fit the particular use case. Although in recent years the NLP community has invested considerable efforts in the generation of resources, most of them are only available for English. However, it is necessary to spend efforts on the development of resources and systems adapted to other languages since, for example, in the specific case of hate speech detection, there are important cultural differences (jargons, slang, idiomatic expressions...) depending on the language or the social group under examination[4, 5]. The general trend in artificial intelligence to data-driven solutions demands a growing volume of data for both training and evaluation. In order to grant greater research focus to data quality and promote data excellence, it is also necessary to work on the construction of specific evaluation sets to improve the quality of training and test data. In addition, it is also necessary to advance in the development of algorithms to build or optimize such datasets, like the creation of "core" sets to deploy augmentation approaches, or more concern in the debugging of text labeling errors.

Finally, the availability of tools to facilitate the task of moderation for human professionals dedicated to the detection of inappropriate language in social media (comments on news in the media, posts in social networks, responses to messages in online platforms...) is becoming increasingly necessary due to the exponential growth in textual interactions of Internet users. The popularization of the use of social media by the vast majority of the population makes it completely unfeasible to moderate comments and messages by strictly manual means performed by humans. The development of tools is required to assist moderators in decision making through early detection of this type of inappropriate behavior on the Web. Having tools that not only indicate the level of toxicity of a text but also identify this type of sexist, xenophobic, homophobic, racist or simply bad-sounding expressions will help us to build a much more inclusive, friendly, social and fair discourse.

We propose this project "MODERATE: MODERAtion of ConTEnts in Social networks using Language Technologies" ((TED2021-130145B-I00)to address these issues. This work has been partially supported by MCIN/AEI/10.13039/501100011033 and by the “European Union NextGenerationEU/PRTR".

The main objective of this project is to study and develop effective solutions based on advanced machine learning and NLP techniques to determine the degree of offensiveness/toxicity/hope speech/hate speech of a text in Spanish and identify the spans related to this type of content. It is proposed to develop a cloud service capable of automatically labeling/classifying texts according to their level of offensiveness or toxicity. In addition, an interface will be designed to detect terms or passages that contain inappropriate language (abusive, offensive, foul language, hate speech...), marking them in such a way as to facilitate their substitution so that the final text eliminates or reduces the degree of offensiveness.

[4] https://www.interior.gob.es/opencms/pdf/archivos-y-documentacion/documentacion-y-publicaciones/publicaciones-descargables/publicaciones-periodicas/informe-sobre-la-violencia-contra-la-mujer/Informe_evolucion_delitos_de_odio_en-Espana_2019_126200207.pdf

[5] https://www.interior.gob.es/opencms/pdf/archivos-y-documentacion/documentacion-y-publicaciones/publicaciones-descargables/publicaciones-periodicas/informe-sobre-la-evolucion-de-los-delitos-de-odio-en-Espana/Informe_evolucion_delitos_odio_Espana_2022_126200207.pdf

The service will work both as an assistant to guide the user who writes a text and avoid possible offensive expressions, as well as for the construction of listening, monitoring and filtering tools in real time on social media (comments, responses, posts...). A prototype will be built with different modules that will be integrated into a single platform that will be freely accessible through an Application Programming Interface (API) and through a demonstrator that will allow the configuration of filters and search parameters, in order to obtain different types of reports.

2. Goals

The main objective of this project is to study and develop effective solutions based on advanced machine learning and NLP techniques to determine the degree of offensiveness/toxicity of a text in Spanish. It is proposed to develop a cloud service capable of automatically labeling/classifying texts according to their level of offensiveness or toxicity. In addition, an interface will be designed to detect terms or passages that contain inappropriate language (abusive, offensive, foul language, hate speech...), marking them in such a way as to facilitate their substitution so that the final text eliminates or reduces the degree of offensiveness. The service will work both as an assistant to guide the user who writes a text and avoid possible offensive expressions, as well as for the construction of listening, monitoring and filtering tools in real time on social media (comments, responses, posts...). A prototype will be built with different modules that will be integrated into a single platform that will be freely accessible through an Application Programming Interface (API) and through a demonstrator that will allow the configuration of filters and search parameters, in order to obtain different types of reports.

To achieve this goal, innovative techniques in NLP and advanced machine learning algorithms will be studied, designed and evaluated, including the generation of language models, application of deep learning and transfer learning techniques, as well as the latest experiences in multitask learning and zero-shot learning that allow the detection of inappropriate behaviors, including offensive language and hate speech, through the integration of sentiment analysis techniques, emotion recognition or toxicity detection, among others. These affective computing elements are proving to be very efficient in the detection of inappropriate language, as they go beyond simple pattern detection, surface learning or lexicon searches, integrating advanced knowledge extracted from large text collections into the trained models.

This global goal can be divided into the following subobjectives:

- Construction and compilation of new tools and resources based on human language technologies to infer, create and utilize knowledge applied to digital content, focusing on the creation of semi-assisted annotators and their application in the annotation process to generate labeled data sets.
- Identification of valid technologies for "listening" the interactions of individuals with their digital and social environment, so these interactions can be further analysed.
- Study, development and implementation of language technologies together with advanced machine learning algorithms focused on the detection of inappropriate behavior on social networks and hope speech.
- Study and development of deep learning algorithms to model different targeted forms of aggressive communication or risky situations, building artificial intelligence solutions to protect citizens.
- Development and implementation to determine the degree of offensiveness/toxicity/hope speech/hate speech of a text in Spanish.

3. Methodology

To achieve these objectives, a methodology and recommendations on search tools in social networks will be defined and established. Filters and search parameters will be defined, and artificial intelligence algorithms based mainly on language technologies (language models, entity recognition, external knowledge integration, author profiling...) and advanced machine learning techniques oriented to neural networks (deep learning, transfer learning, multitask learning, zero-shot learning) will be proposed, although classic techniques such as SVM or logistic regression will also be evaluated. Existing and new or adapted linguistic resources that can contribute to the discovery of aggressive language patterns will also be taken into account. We will analyze which social networks and social media are the most suitable for monitoring and how to approach the problem in each of them.

The first step will be an in-depth study of the state of the art as well as an analysis of the problem. We will also decide the scenarios and social media to work on, establishing a working methodology for each of them. After the analysis and establishment of the methodology, we will start with the design and implementation of the platform to help to detect hate speech and offensive language. A prototype with the integration of the different modules will be freely accessible, allowing the configuration of filters and search parameters, in order to obtain different types of reports. During the development, various learning algorithms, language models and

linguistic resources will be tested and evaluated in order to refine the final result and adjust the application for the task of analysis and detection of offensive language discourse. For each of the selected domains and scenarios, the following activities will be carried out:

1. Data collection: Good practices will be applied for the retrieval, processing and storage of big data. Since we will be working with textual information obtained from social networks, it is important to identify those technologies that are valid for "listening" and storing the interactions of individuals with their digital and social environment.
2. Design of techniques and tools based on NLP: Development of algorithms based on neural networks to model different specific forms of communication, building artificial intelligence solutions that will act as early detection systems. These algorithms will be trained from available labeled datasets. Furthermore, once these algorithms are implemented and start collecting interactions, the algorithms will be fine-tuned by re-training with the captured and classified information to adapt to the variability and flexibility of language over time: slang, ill-formed expressions, typos, grammar that occur very often in social networks and other interpersonal communications.
3. Integration of technologies: These trained models will be incorporated into a web application that will act as a monitor of offensive language messages in social networks, so that, for each message retrieved, its content will be automatically analyzed to detect whether it contains hate speech, offensive language, toxicity... and the result will be visualized through different graphs. The user will have the possibility to generate a report and analyze the content of the data based on filters.
4. Evaluation of automatic systems: An essential step is the evaluation of the developed algorithms to estimate the accuracy and coverage that these algorithms could have in a real scenario, i.e. when monitoring the Web to detect this type of inappropriate behavior.

The different activities described in the previous steps can be visualized in Figure 1.

4. Scientific and Technical Impact

The MODERATES Project focuses on a series of scientific challenges in the field of Human Language Technologies (HLT) which must be faced with the experience and techniques developed by the research team involved in the project. Firstly, there exists an important lack of resources, even more when dealing with the Spanish language. In the project we will focus on generating data and developing methods, techniques and tools to secure our interactions in the digital society. New architectures for artificial intelligence represent the main approach to address social media monitoring, but further questions remain, as new linguistic resources, specific and adapted language models and integration solutions are needed to reach the objectives proposed by this project.

On the other hand, the impact of the COVID-19 pandemic has leveraged our dependency to the digital channels of communication such as social media and social networks, increasing the potential negative effect on the population, especially among vulnerable communities such as children and teenagers, women, LGTBI, immigrants, religious and cultural communities.

For all these reasons it is expected that the project will have a scientific impact (both nationally and internationally) in different fields, such as the creation and use of effective linguistic, semantic and computational resources for different languages, especially in Spanish; the development and integration of specific methods and tools for the retrieval, extraction, automatic analysis, summarization and representation in digital entities of information coming from different textual genres as well as the improvement or creation of resources and components such as crawlers, sentiment and opinion analysis systems and integration tools, which will have a strong impact on the scientific community and society. It is also expected to have transferable results in the medium term, working with real practical cases.

5. Social and Economic Impact

This proposal focuses on a significant strategic societal challenge given that the detection and prevention of inappropriate behavior on the Internet is of the greatest importance for the contribution to the goal of smart, sustainable and inclusive growth within Europe in accordance with European, national and regional policy-making.

Digital media have become a space where hoaxes, hate speech or abusive behavior proliferate, among other contents that directly and negatively harm the users of this space in particular and society in general. Thus, this project will offer the modeling of the behavior of digital content, being able to contribute, in the detection, mitigation and prevention of harmful digital content, in pursuit of a sanitation of social media on the Internet helping to ensure a respectful, safe and reliable communication environment.

The project will have an important social impact from the point of view of digital content modeling. This modeling will provide a framework for specialists to develop

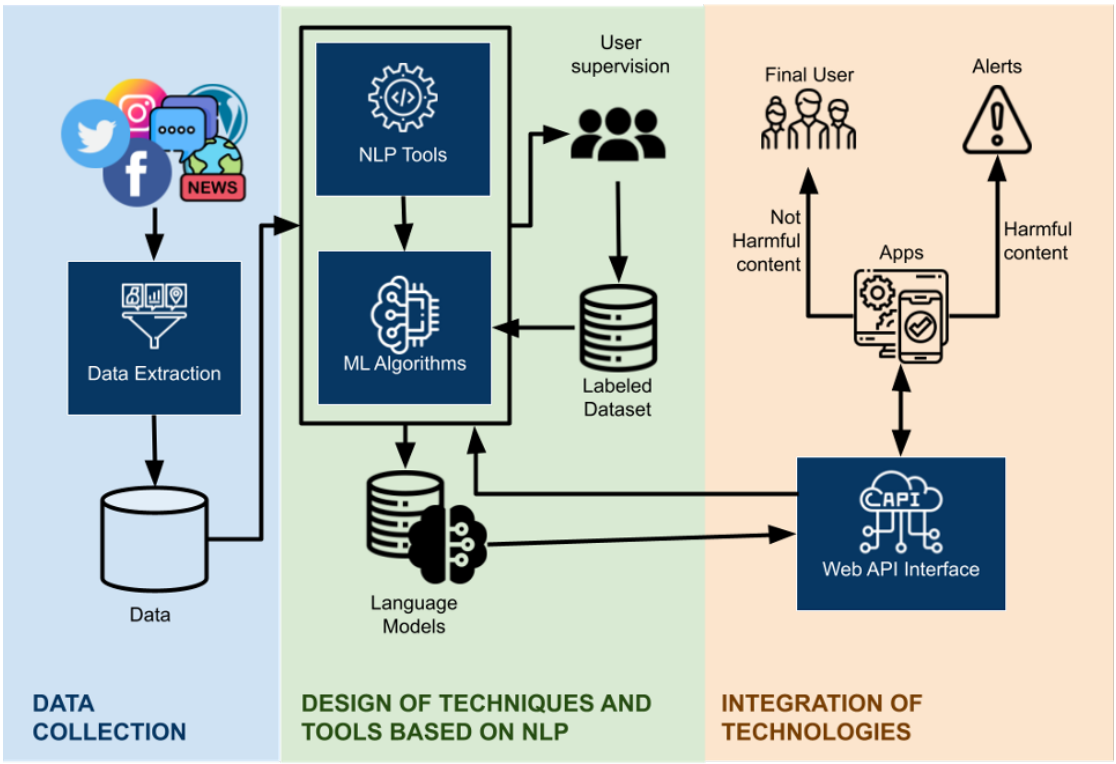

Figure 1: Detailed activities of the research and integration processes

and implement information systems to address negative social phenomena, protect society from dangers that may be posed to citizens, etc. Also, positive digital content can improve digital literacy and people's e-safety skills and awareness to prevent the risks of harmful content on the Internet, and thus improve users' experience on the Internet. Given that service providers have legal responsibility for content, having tools for monitoring illegal hate speech would help to comply with the protocols imposed by legislation .

In summary, a direct social impact is to reduce the harmful effects that inappropriate content causes on social networks, especially among the most vulnerable sectors of society: young people and children, people with disabilities and people susceptible to harassment (women, LGTBI, immigrants, religious and cultural communities).

In terms of economic impact, several sectors could benefit such as communication media, tourism, fake-news detection, digital security, hate speech detection, constructive discourse promotion or self-learning. On the other hand, big tech companies could be interested in this technology, also providing a great economic impact on the transfer of this knowledge as mentioned in Section 2 of the project justification. Companies like Facebook or Twitter spend huge amounts of money developing automatic models for detecting inappropriate content. Major social media companies spend huge amounts of money developing automatic models for detecting inappropriate content. The estimated investment in content moderation will rise from 5 billion today to almost 12 billion in the next five years. Due to problems with human moderation, companies already make use of artificial intelligence for automatic moderation in 60% of the content. Having platforms and services such as the one proposed by the project would have a series of clear impacts on content moderation: 1) Increase the performance of manual moderation, 2) Reduce the costs of developing and integrating automatic detection solutions, and 3) Promote the moderation of content in Spanish, a language that already accounts for 8% of all communications on social networks, with a constant increase each year (reaching 15.6% in the most widespread network, Facebook).

Despite the clear interest of such large corporations, numerous startups and technological mid-size companies are offering automatic tools for textual analysis in a wide range of sectors such as press, social media, e-commerce...

Content moderation is needed by many online services (anyone where clients of users are able to publish comments). The technology and solutions released by the project could help local and international companies in providing more robust and performance products, empowering Spanish oriented products and improving competitiveness for such as mid-size companies.

Acknowledgments

This work has been partially supported by project MODERATES (TED2021-130145B-I00) funded by MCIN/AEI/10.13039/501100011033 and by the “European Union NextGenerationEU/PRTR"

References

[1] A. S. Parihar, S. Thapa, S. Mishra, Hate speech detection using natural language processing: Applications and challenges, in: 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), IEEE, 2021, pp. 1302–1308.

[2] 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.

[3] S. M. Jiménez-Zafra, M. Á. Garcia-Cumbreras, D. García-Baena, J. A. Garcia-Díaz, B. R. Chakravarthi, R. Valencia-García, L. A. Ureña-López, Overview of hope at iberlef 2023: Multilingual hope speech detection, Procesamiento del Lenguaje Natural 71 (2023) 371–381.

[4] F. M. P. Del Arco, A. B. P. Portillo, P. L. Úbeda, B. Gil, M.-T. Martín-Valdivia, Share: A lexicon of harmful expressions by spanish speakers, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 1307–1316.

[5] F. M. P. Del Arco, A. Montejo-Ráez, L. A. U. Lopez, 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), 2021, pp. 1096–1108.