A Social Monitor for Detecting Inappropriate Behavior

Monitor Social para Detectar Comportamientos Inapropiados

José Alberto Mesa Murgado, Flor Miriam Plaza-del-Arco,
Pilar López-Úbeda, M. Teresa Martín-Valdivia
Departamento de Informática, CEATIC, Universidad de Jaén, España
{jmurgado, fmplaza, plubeda, maite}@ujaen.es

Abstract: In this paper we present a prototype of a social monitor aimed at detecting inappropriate behavior in social networks applying Machine Learning (ML) solutions. This monitor is able to retrieve textual data from two of the current main social networks (YouTube and Twitter) and analyze them using ML algorithms based systems trained on different scenarios. In particular, those that have emerged in recent years in the Natural Language Processing (NLP) community and that are related to social and mental health issues, such as the detection of offensive language, cyberbullying, fake news and the identification of mental disorders.
Keywords: Social media Mining, Machine Learning, Natural Language Processing.

Resumen: En este artículo presentamos un prototipo de monitor social destinado a detectar comportamientos inapropiados en las redes sociales, aplicando soluciones de aprendizaje automático. Este monitor es capaz de recuperar datos textuales de dos de las principales redes sociales hoy en día (YouTube y Twitter) y analizar esta información haciendo uso de sistemas basados en aprendizaje automático entrenados en diferentes escenarios. En particular, aquellos que han surgido en los últimos años en la comunidad del Procesamiento del Lenguaje Natural (PLN) y que están relacionados con importantes retos sociales, como por ejemplo la detección de lenguaje ofensivo, ciberacoso, noticias falsas y la identificación de trastornos mentales.
Palabras clave: Minería de datos en redes sociales, Aprendizaje Automático, Procesamiento del Lenguaje Natural.

1 *Introduction and Motivation*

The emergence of the Web 2.0 has completely changed the way people communicate and interact. The most popular social networks such as YouTube, Twitter or WhatsApp have more than 2,000 million registered users. Every second, on average, over 6,000 tweets are published on Twitter, which results in over 500 million tweets per day[1].

Given the large amount of data accessible via the Web, different studies in the field of Natural Language Processing (NLP) have emerged seeking to offer solutions to society in different areas such as psychology to identify people moods (Plaza del Arco et al., 2020) mental disorders (López-Úbeda et al., 2019) (López-Úbeda et al., 2021b), marketing to analyze product reviews to boost sales (Rambocas and Pacheco, 2018), and sociology to detect hate speech (Plaza-del Arco et al., 2021) or constructive news (López-Ubeda et al., 2021a).

On the one hand, every day, many current events go viral on social media, usually related to political issues, celebrities, video games, fashion or diseases. On the other hand, although we find many studies applying ML solutions to analyze this data, the availability of tools that integrate these systems to be used in real scenarios is scarce.

In this paper we present a prototype of a social monitor aimed at detecting inappropriate behavior through the following functionalities: *i*) collecting data from two of the main social network: Twitter and YouTube; *ii*) integrating ML systems trained on different tasks, and *iii*) analyzing the data col-

[1] https://www.internetlivestats.com/twitter-statistics/

lected using NLP solutions. With this prototype, ML-based systems can be applied to real scenarios that test their performance. It should be remarked that the monitor is especially aimed at integrating NLP solutions, which have arisen over the last few years within the scientific community, to tackle social challenges in social networks such as cyberbullying detection, fake news and mental disorders identification.

The rest of the paper is structured as follows: In Section 2 we present the description of the tool in which we will detail information related to backend and frontend development. The real scenarios where the tool will be implemented are explained in Section 3. Finally, Section 4 presents conclusions and future work.

2 System Description

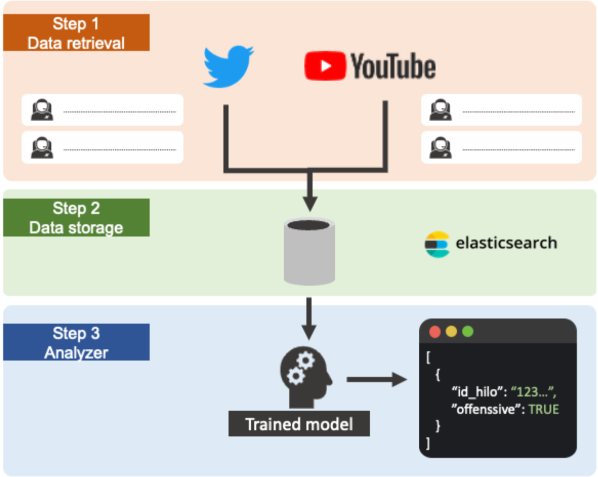
The system is a social monitor designed to retrieve posts from two of the current main social networks: Twitter and YouTube. The purpose of collecting such posts is to analyze them using models trained on well known tasks of the NLP community such as text classification and Named Entity Recognition.

This tool is implemented using Python's FastAPI[2] framework to establish the connection between users, database and the official social networks APIs through POST and GET HTTP requests. This framework has been chosen due to its advantages: high performance, detailed documentation, strong relationship with API development standards such as JSON schemas, and finally, for the ease of learning and developing using it.

The architecture of this tool is displayed in Figure 1. First step consists of extracting posts from different social networks, specifically Twitter and YouTube. Subsequently, these posts are stored in a database (step 2). Finally, with the idea of analyzing the information extracted from these social networks, posts are evaluated using pre-trained ML systems (step 3).

2.1 Backend

This side of the application is responsible of extracting information from different social networks (Section 2.1.1), storing data (Section 2.1.2) and integrating different NLP models that have been previously trained on a specific corpora for a specific task(Section 2.1.3).



```
[
  {
    "id_hilo": "123...",
    "offenssive": TRUE
  }
]
```

Figure 1: Social Monitor Architecture.

2.1.1 Data Retrieval

Twitter posts and thread. To retrieve posts from Twitter, we use its official API[3]. It limits the number of tweets that can be requested to 25,000 monthly under the standard and academic license[4] (the enterprise license extends this limitation up to 5 million tweets per month). The social monitor allows the possibility of receiving tweets extracted from Twitter regarding:

1. An user timeline and the associated responses to those tweets.
2. Contained in a given hashtag.
3. Nearby a given location defined in terms of latitude and longitude.

Specifically, we look for tweets written in Spanish although this restriction could be updated to fit other requirements. Tweets retrieved from this request are stored inside our Elasticsearch index.

YouTube video comments. To retrieve comments from YouTube videos Google provides an official API[5]. To request this data from this API we use Google's official Python library[6] along with our credentials. The number of elements to retrieve is restricted

[2] `https://fastapi.tiangolo.com/`

[3] `https://developer.twitter.com/en/docs/twitter-api`

[4] `https://developer.twitter.com/en/docs/twitter-api/rate-limits`

[5] `https://developers.google.com/youtube/v3`

[6] `https://github.com/googleapis/google-api-python-client`

up to 10,000 per day. Comments to YouTube videos can be retrieved regarding:

1. **Channel or user identifier** to focus on, the API deepens the search in its latest video releases.
2. **Hashtag or term** present either as part of the video's title or its associated description.
3. **Location in terms of latitude and longitude** from which to look for nearby videos.

Besides these parameters, requests must declare how many videos the search for comments should be performed on and how many of these comments are to be retrieved per video at most. YouTube data API does not discriminate comments by language and this should be performed by a third party agent such as the spaCy library for Python [7].

2.1.2 Data Storage

Elasticsearch[8] search engine has been used to save the posts because it allows us to store, search and analyze vast volumes of structured and non-structured data within a fast response time.

Every post extracted from Twitter or YouTube is stored in our Elasticsearch index containing the folowing data:

- **Post identifier**, it is the identifier associated to the user post.
- **Thread identifier**, given the case that the item is a response to another post.
- **Textual comment** written by the user.
- **Date** on which the post is published.
- **Name of the social network** where the post is retrieved.

2.1.3 Integrating Machine Learning Models

The social monitor also allows the integration of different models based on ML to analyze and evaluate the information stored in Elasticsearch. These systems are previously trained on different NLP tasks.

For this purpose, the user could select the model to be used and the information to be analyzed, for example.

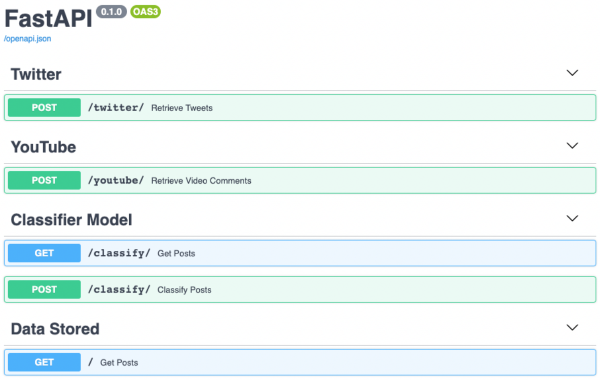
Moreover, the tool offers high flexibility and adaptability as it allows the incorporation of different systems, trained by using several methods for different tasks and languages. Specifically, this monitor relies on traditional ML systems such as Support Vector Machine (SVM) and other state-of-the-art methods based on Transformer models such as BERT.

2.2 Frontend

An automatic visual and interactive interface has been generated using Swagger UI[9]. The functionalities offered by the tool are described below:

1. Collect posts from the different social networks (Twitter and YouTube) and store them in the Elasticsearch database.
2. Retrieve previously stored data to produce a graphical representation of the information.
3. Perform an analysis to analyze the stored comments using the NLP models integrated into the tool.

Figure 2 provides a screenshot of the frontend of the social monitor to enhance the above functionalities.



FastAPI 0.1.0 OAS3

/openapi.json

Twitter

POST /twitter/ Retrieve Tweets

YouTube

POST /youtube/ Retrieve Video Comments

Classifier Model

GET /classify/ Get Posts

POST /classify/ Classify Posts

Data Stored

GET / Get Posts

Figure 2: Social monitor frontend.

3 *Real Scenarios*

Since the data extracted from the different social networks are stored in a database, we have integrated different systems based on ML to the tool in order to evaluate each post.

So far the system has been trained to detect inappropriate behavior on social media

[7] https://spacy.io/

[8] https://www.elastic.co/es/elasticsearch/

[9] https://swagger.io/tools/swagger-ui/

and address two challenging tasks for the NLP community. The first one is anorexia detection, identifying whether a post contains information related to this eating disorder. The second, offensive language identification or recognizing whether a tweet contains hurtful, derogatory, or obscene terms. These models are based on ML systems including SVM (Noble, 2006) and state-of-the-art Transformers models such as BERT (Devlin et al., 2018). Figure 3 shows an output example of the monitor after analyzing a tweet using the offensive language detection trained model.

```
"offensiveness": "TRUE",
"id_thread": "1234",
"id_post": "1254",
"comment": "Como @user puede ser
tan tonto, es que no me lo explico"
(How can @user be so dumb?
I don't get it),
"time": "2021-01-22T12:24:18",
"source": "Twitter"
```

Figure 3: Example of the social monitor output after analyzing a retrieved tweet.

4 Conclusions and Future Work

In this paper we present a prototype of a social monitor aimed at detecting inappropriate behavior in social media: Twitter and Youtube. For this purpose, the tool includes a database where the user can store the posts from the social networks to later analyze them with different previously trained NLP systems. Specifically, the tool integrates two systems based on document classification to identify anorexia and offensive language although it is implemented in a way that allows its use on other tasks.

For future work, we plan to incorporate visual analytics using Kibana[10] to display statistics associated with the stored data. In addition, we will integrate NLP trained models related to different topics of interest for the scientific community using data retrieved from social networks. In this way, the same database will be useful for different purposes.

Acknowledgements

This work has been partially supported by a grant from Fondo Europeo de Desarrollo Regional (FEDER), LIVING-LANG project [RTI2018-094653-B-C21], and the Ministry of Science, Innovation and Universities (scholarship [FPI-PRE2019-089310]) from the Spanish Government.

References

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

López-Ubeda, P., F. M. Plaza-del Arco, M. C. Díaz-Galiano, and M. T. Martín-Valdivia. 2021a. Necos: An annotated corpus to identify constructive news comments in spanish. *Procesamiento del Lenguaje Natural*, 66:41–51.

López-Úbeda, P., F. M. Plaza-del Arco, M. C. Díaz-Galiano, and M. T. Martín-Valdivia. 2021b. How successful is transfer learning for detecting anorexia on social media? *Applied Sciences*, 11(4):1838.

López-Úbeda, P., F. M. Plaza del Arco, M. C. Díaz-Galiano, L. A. Ureña-López, and M. T. Martín-Valdivia. 2019. Detecting anorexia in spanish tweets. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 655–663.

Noble, W. S. 2006. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567.

Plaza-del Arco, F. M., M. D. Molina-González, , and M. T. Martín-Valdivia. 2021. Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, 166:114120.

Plaza del Arco, F. M., C. Strapparava, L. A. Urena Lopez, and M. Martin. 2020. Emo-Event: A multilingual emotion corpus based on different events. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association.

Rambocas, M. and B. G. Pacheco. 2018. Online sentiment analysis in marketing research: a review. *Journal of Research in Interactive Marketing*.

[10] https://www.elastic.co/es/kibana